DETECTION AND EXPLANATION OF STATISTICAL DIFFERENCES ACROSS A PAIR OF GROUPS

by

Yuriy Sverchkov

B.S., University of Maryland, Baltimore County, 2008

M.S., University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yuriy Sverchkov

It was defended on

July 16, 2014

and approved by

Gregory F. Cooper, Department of Biomedical Informatics

Giles Clermont, Deparment of Critical Care Medicine

Milos Hauskrecht, Department of Computer Science

Shyam Visweswaran, Department of Biomedical Informatics

Dissertation Director: Gregory F. Cooper, Department of Biomedical Informatics

DETECTION AND EXPLANATION OF STATISTICAL DIFFERENCES ACROSS A PAIR OF GROUPS

Yuriy Sverchkov, PhD

University of Pittsburgh, 2014

The task of explaining differences across groups is a task that people encounter often, not only in the research environment, but also in less formal settings. Existing statistical tools designed specifically for discovering and understanding differences are limited. The methods developed in this dissertation provide such tools and help understand what properties such tools should have to be successful and to motivate further development of new approaches to discovering and understanding differences.

This dissertation presents a novel approach to comparing groups of data points. The process of comparing groups of data is divided into multiple stages: The learning of *maximum a posteriori* models for the data in each group, the identification of statistical differences between model parameters, the construction of a single model that captures those differences, and finally, the explanation of inferences of differences in marginal distributions in the form of an account of clinically significant contributions of elemental model differences to the marginal difference. A general framework for the process, applicable to a broad range of model types, is presented. This dissertation focuses on applying this framework to Bayesian networks over multinomial variables.

To evaluate model learning and the detection of parameter differences an empirical evaluation of methods for identifying statistically significant differences and clinically significant differences is performed. To evaluate the generated explanations of how differences in the models account for the differences in probabilities computed from those models, case studies with real clinical data are presented, and the findings generated by explanations are

discussed. An interactive prototype that allows a user to navigate through such an explanation is presented, and ideas are discussed for further development of data analysis tools for comparing groups of data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Groups of data collected on a given set of variables may reflect different underlying distributions. It is often helpful to determine if those distributions differ, and if so, explain how they differ. For example, a pair of variables may show correlation in one group, but independence in another group; or we may see an association between a pair of variables preserved across groups, but observe a difference in marginal distributions. There are other possibilities, and when considering data over many variables, there are many more relationships to consider. Cases where the statistical comparison of groups of data is of interest arise in widely varying applications, including clinical research, quality assurance, comparative effectiveness research, and many others. An example of a quality assurance scenario is one where we observe that two intensive care units (ICUs) of the same type experience notably different readmission rates. A quality assurance officer might want to discover whether particular differences in the operation of the ICUs contribute to the difference in readmission rates. An example in the clinical research setting would be that of a cohort study that investigates the association of early dialysis and the mortality of renal disease patients. In such a study a clinical researcher would want to identify those circumstances under which an effect is present, and the extent to which the effect is positive or negative. Another example from clinical research is that of a randomized controlled trial, where we want to identify patterns in the response to a treatment. A biologist may seek to perform an exploratory analysis of the differences between measurements of cancer cells and measurements of healthy tissue. All of these scenarios share a common structure: there are two groups for which data are available and the goals are to determine if the groups differ from each other, and if so, how they differ.

1.1 MOTIVATION

This dissertation develops and evaluates a method for analyzing data on two groups that share a common set of variables in order to determine if the groups differ in their multivariate distributions over those variables, and if so, how. The dissertation focuses on the problem of comparing a pair of groups.

Many questions in research and everyday life can be framed as questions of difference explanation. Interestingly enough, the statistical tools typically available to researchers are not focused on this task, but rather are focused on either predicting an outcome, or classifying cases, or clustering cases. Tests that do aim at difference identification are not typically good as tools for explanation. Tests of statistical differences are generally either univariate (chi-square, t-test, Kolmogorov-Smirnov test), or when they are multivariate (Hotelling's T-squared test, Kullback-Leibler divergence), they cannot relate the difference measure to how variables individually contribute to that overall measure.

Take a simple example of explaining to a child how an elephant is different from a mouse. A collection of univariate difference tests might detect that elephants are bigger, eat more, drink more, move less swiftly, etc. A multivariate test might say that indeed, an elephant is different from a mouse, but would not really give a reason as to why. A good classifier would say that if it's an animal that has a trunk, it's probably an elephant.

None of these seem satisfactory. What explanation would we want to see? Something along the lines of "elephants bigger mammals than mice, and bigger mammals tend to eat more, drink more, move less swiftly because of their size. Elephants also have trunks, while mice have pointed snouts."

The key to this explanation is that it uses a model that captures the differences and similarities between mice and elephants (a portion of the model is dedicated to correlating a mammal's size to diet and biophysical constraints, a relationship that is applicable to both elephants and mice, while another portion captures the differences). We see a phenomenon that a long list of univariate differences can be explained away by using the model and tracing many differences to a difference in one key observable variable—size. Some differences (trunk/snout) still need to be pointed out independently. We also see that what may be

sufficient for a classifier is not sufficient to providing a full account of differences of interest.

This observation suggests an approach to difference explanation that focuses on explicitly modeling each of the groups we compare as well as their commonalities. This dissertation focuses on identifying and explaining differences in groups of data, where a group of data is defined as a collection of points of data, such as a collection of medical records, where each data point, or record, consists of a set of measurements about, for example, a patient that belongs to one of the groups we are comparing. The probabilistic approach taken in this dissertation assumes that the records in each group are independent and identically distributed samples that come from some joint probability distribution common for that group, and that the individual measurements are the random variables governed by that probability distribution. The general task of finding differences between two groups of records is then framed as that of finding the differences between the probability distributions. Since we rarely (if ever) know what the precise probability distribution for data is, I take a Bayesian approach of estimating the probability distribution of each group by the expected value over a distribution of posterior distributions based on the data and prior knowledge, which may be informative or non-informative. A Bayesian approach also allows us to evaluate the statistical significance of estimates that are based on a distribution of distributions. There are various techniques for modeling probability distributions, and the sorts of models that produce informative explanations are typically ones that represent the full joint probability distribution in terms of *local* probability distributions, that is, ones that involve small subsets of the full set of variables. For such models, an analysis of differences and their effects on variables of interest consists of identifying differences in local distributions and explaining how the effects of these differences combine to account for differences in the distributions of the variables of interest.

The general hypothesis of this thesis is as follows: *One can systematically produce explanations that are more revealing and insightful than those obtained from traditional methods by approaching the problem of comparing a pair of groups as that of identifying significant local distributional differences between two multivariate distribution estimates for those groups and explaining their effects on variables of interest.*

1.2 CONTRIBUTIONS OF THIS THESIS

The main aim of this dissertation is to help address the general problem of detecting and explaining differences between groups of records in terms of probability distributions. The two major components to addressing the general problem and the general hypothesis stated above are: the identification of the differences across the groups in the context of a model (Chapters 4 and 5), and the utilization of this model to explain the differences in the data and their interrelations (Chapters 6 and 7).

I describe broadly a framework for the identification of distributional differences across groups in the context of a model that codifies those differences. George Box famously said that "all models are wrong, but some are useful" (Box and Draper, 1987). The commitment to a model that is 'wrong' by nature of being a model, then, introduces the risk that inferences drawn from the model would not necessarily match direct observation. One might wonder, then, especially with the current increasing availability of large datasets and the computing power to query the data directly, whether querying the data for counts or observed proportions is not a better method for identifying patterns of differences. In spite of this concern, the use of a model is essential to producing an explanation that points to the differences in the mechanisms which govern the data. If we view the data in the two groups that we compare as produced by two generative models, the difference detection methods here aim to find the similarities and differences between the processes that produce the data, and relate the differences observed in the generated data to the differences that we find between the processes. I do not believe that such an explanation is possible without the construction of a model that captures the differences. Moreover, the risk of committing to a model that has such a mismatch with the data can be mitigated by considering ensembles of models, or Bayesian averaging over a space of models; Section 8.2.11 discusses the possibility of applying these approaches in future work.

The framework for identifying differences in distributions that I describe can be applied using a variety of probabilistic models of data, including Bayesian networks (BNs), Markov random fields, Bayesian hierarchical models, probabilistic rule-based systems, and others. In this dissertation, the focus is on applying this framework to BNs of random variables

that have multinomial distributions. The difference detection approaches developed here can more broadly apply to any model that meets two criteria. The first is that the models be described in terms of local parameters that are independent in their prior distributions. In BNs, this corresponds to local and global parameter independence (Heckerman et al., 1995). The second is that it be possible to match parameters across models for the different groups. In BNs, this is easy to accomplish since parameters correspond to the conditional distribution that a variable takes given the values of other variables. Given a model that has such parameter distributions, the difference detection and explanation task is to identify differences in those matching local parameters and explain their effect on the variables and inferences of interest.

There are two types of criteria used to determine whether parameters being compared are different: statistical significance and clinical significance. Statistical significance is a measure of the evidence that the difference observed in data is not due to random chance (Coolidge, 2012). "Clinical significance refers to the practical or applied value or importance of the effect of an intervention" (Kazdin, 1999). In this dissertation, the term "clinical significance" is used more broadly to mean that the effect size is large. In statistics, the *effect size* is a quantitative measure of the strength of a phenomenon (Kelley and Preacher, 2012). Section 4.4 details the particular measures of effect size used in this dissertation. They are used to evaluate the deviation of an observed quantity from its expected value under a null hypothesis. The concepts of clinical and statistical significance are generally applicable to various types of measurements.

To understand the distinction between these two types of significance in the context of this dissertation, consider and example with a multinomial variable X_i, where the measurement of interest is the difference in the proportion of cases for which X_i takes a value x_{ik} across the two groups. This dissertation focuses on distinguishing the presence of differences from the absence of differences. Therefore, one measure of effect size in which we are interested is the deviation of the difference in probabilities from zero. Suppose that in one group $P(x_{ik}) = 0.100$ and in the other, $P(x_{ik}) = 0.101$. It is entirely possible that with a large amount of data, we may have enough evidence to conclude that the difference would be statistically significant, that is, not likely to be due to random variation in the data. This

difference, however, would probably not be considered to be clinically significant because the two probabilities are too close for the difference to be of importance in practical purposes.

In this dissertation, statistical significance is used to determine relationships between variables in the two groups. Statistical significance is used to determine which variables are dependent on which, and whether model parameters are different across the two groups compared. Clinical significance is used to further evaluate differences that were found to be statistically significant.

Once a model of the data is constructed and the local parameter differences across the groups have been identified (via statistical significance testing), explanation can be performed. The explanation methods in this thesis are tailored for BNs with multinomial variables, and while they may be extended to other models that are semantically similar to BNs, ultimately, the semantics of a model will direct how an explanation regarding the information it captured is performed. The explanation process consists of identifying clinically significant marginal differences in variable distributions and tracing the difference to its sources in the parametric differences in the model definition. Tests of clinical significance are used to determine which elements are significant enough to include in the explanation.

To summarize, the specific contributions are as follows: the formal framing of the group-difference problem in terms of model parameter difference identification, the evaluation of methods for detecting statistically significant parameter differences and clinically significant parameter differences, and the development and evaluation of methods for generating a comprehensive explanation of differences that trace a difference of interest to fundamental parameter differences in a model of the data.

1.3 DISSERTATION OUTLINE

The remainder of this dissertation is structured as follows. Chapter 2 presents background concepts that are integral to the research and summarizes data and notation used. Chapter 3 reviews related research. Chapter 4 presents a conceptual framework for finding clinically and statistically significant differences in data by leveraging model learning. I present two

applications of this framework using BN models. Chapter 5 evaluates the detection methods presented in Chapter 4. Chapter 6 presents three methods for generating explanations given a model of the data and the parameter differences detected with respect to that model. The approaches are presented in increasing order of complexity, with each building on the previous ones. Chapter 7 describes the design and operation of a prototype interactive system for difference detection and explanation, which implements the methods of Chapter 6. A discussion of the prototype's strengths and weaknesses and ideas for future development concludes the chapter. Finally, Chapter 8 summarizes the findings and contributions of the dissertation and discusses possible directions for future research.

2.0 BACKGROUND

2.1 SUMMARY OF NOTATION

This section serves as a brief summary of all notation used in this document. Most of the notation is also defined as it is introduced in the text.

2.1.1 Symbols and operators

- Logical operators $\wedge, \vee, \oplus, \neg$: and, or, exclusive or, and not, respectively.
- $\neg$ is overloaded to also flip the value of binary variables.
- Set operators $\cup, \cap, \setminus$: union, intersection, and set subtraction, respectively. Set operation are also sometimes applied to vectors when the vectors are treated like sets (element order does not matter in those cases).
- $\cup$ and $\cap$ are overloaded as symbols, usually in subscripts and superscripts, as indicators of union/intersection-like variants e.g. $\mathcal{D}_{\cup}$ indicates the concatenation of data groups, hence $\mathcal{D}_{\cup}$ is in that sense "union-like."
- Probability theory operators $\sim, \perp$: "distributed as" and "independent of," respectively.
- Probability operator $P(\cdot)$ is the probability of an event or the probability distribution of a collection of events that correspond to the various values taken by random variables in the expression on which the operator operates.
- Expectation operator $E_R[f(R)]$, for a random variable R represents the expectation of $f(R)$ that is obtained by averaging over the distribution of R.

2.1.2 Variable naming conventions

- Uppercase Latin or Greek letters are random variables. Exceptions: H, J, K, N are integers (see below).
- Bold symbols are vectors (ordered collections). Exception: Π, π, are not bold, but they do indicate a collection.
- The above combine, for example: $\mathbf{X}$ is a vector of random variables $(X_1, \ldots, X_n)$
- Lowercase letters whose uppercase variant is a random variable indicate a particular assignment of those variables to values. Examples: θ_{ijk} is a particular value that Θ_{ijk} takes, $\boldsymbol{\theta}_{ij}$ is a vector representing particular value-assignment of the random vector $\boldsymbol{\Theta}_{ij}$, π_{ij} is the j-th assignment of Π_i.

2.1.3 Variables with specific meanings

- $\mathbb{N}$ is the set of natural numbers.
- $\mathcal{D}$ represents data. The terms data, a group of data, and data-set all refer to a collection of records (or data-points) over a set of random variables. Each record is an assignment of the full set of random variables to values.
- Specifically, $\mathcal{D}_1$ and $\mathcal{D}_2$ represent the two groups of data to be compared, and $\mathcal{D}_\cup$ is their concatenation.
- $\mathcal{M}$ represents models, which, depending on context, means wither the abstract idea of a model, or specifically, a Bayesian network structure.
- Z is a binary random variable taking the values $\{1, 2\}$, it is called the group-indicator variable, and its value for a particular record is the group to which the record belongs, i.e. $Z = 1$ for all records in $\mathcal{D}_1$ and $Z = 2$ for all records in $\mathcal{D}_2$.
- $\mathbf{X}$ is the collection of variables in a data-set, not including Z.
- n is the size of $\mathbf{X}$.
- $i \in \{1, \ldots, n\}$ is the index of a variable X_i in the vector $\mathbf{X}$.
- K_i is the number of values variable X_i takes.
- $k \in \{1, \ldots, K_i\}$ is the index of an assignment of a variable X_i to a value X_{ik}
- Π_i is the vector of variables that are the parents of X_i in a particular graph structure.

- J_i is the number of possible assignments of Π_i to values.
- $j \in \{1, \ldots, J_i\}$ is the index of an assignment of Π_i to a value vector π_{ij}.
- θ_{ijk} is a parameter in a Bayesian network representing the probability $P(x_{ik}|\pi_{ij})$.
- $\boldsymbol{\theta}_{ij}$ is the vector $(\theta_{ij1}, \ldots, \theta_{ijK_i})$.
- $\boldsymbol{\theta}_i$ is the vector $(\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{iJ_i})$.
- $\boldsymbol{\theta}$ is the vector $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$.
- N_{ijk} is the number of points in a data set for which $X_i = x_{ik}$ and $\Pi_i = \pi_{ij}$.
- $\mathbf{N}_{ij}$ is the vector $(N_{ij1}, \ldots, N_{ijK_i})$.
- $\mathbf{N}_i$ is the vector $(\mathbf{N}_{i1}, \ldots, \mathbf{N}_{iJ_i})$.
- $N_{ij\cdot} = \sum_{k=1}^{K_i} N_{ijk}$.
- $N = \sum_{i=1}^{n} \sum_{j=1}^{J_i} N_{ij\cdot}$, the number of data points in a data set.
- t is the index of a parent Y_t of X_i in Π_i.
- H_i and η are analogous to J_i and j, see Section 4.3 for details.

2.2 BAYESIAN NETWORKS

A BN over a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ consists of a directed acyclic graph (DAG) and a collection of conditional probability distributions. In the DAG, also called the BN structure, each node represents a variable X_i.

The DAG defines an ancestry structure between nodes: a node has outgoing arcs to its children, and incoming arcs from its parents; a parent is an ancestor, and a parent of an ancestor is an ancestor; X_a is a descendant of X_b iff X_b is an ancestor of X_a; the parent set of a node with no incoming arcs is the empty set; an undirected path through the network is one that follows arcs while ignoring their direction; an undirected loop is an undirected path that starts and ends at the same node, passes through at least one other node, and does not cross itself.

Figure 1 is an example of a BN structure that illustrates these concepts. The figure is showing the relationships between various nodes and X_5 as well as an example of a loop and an undirected graph: X_3 and X_4 are the parents of X_5; X_1, X_2, X_3, and X_4 are the ancestors

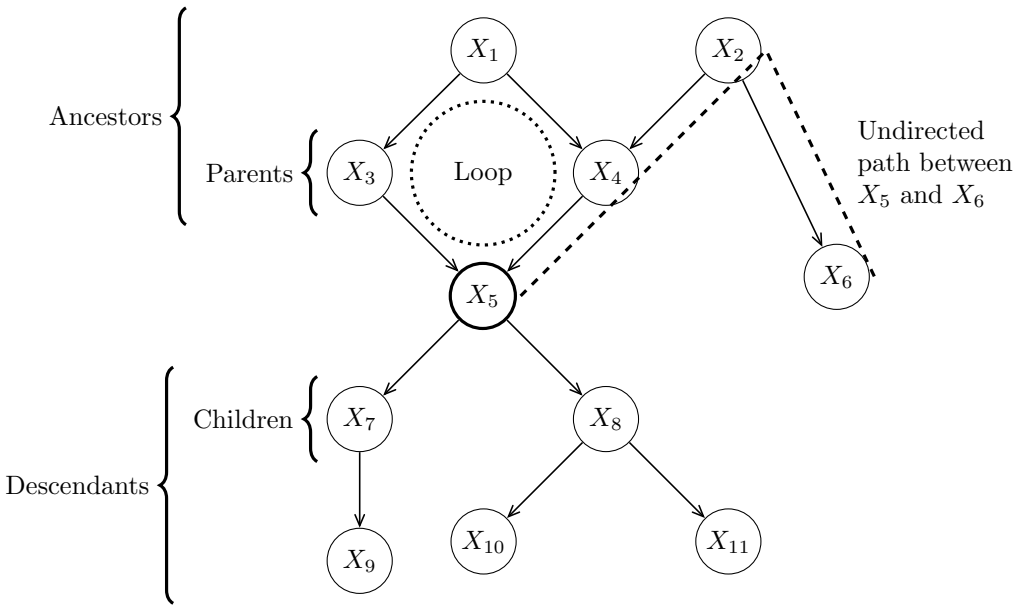

Figure 1: An example illustrating the anatomy of a BN and graph concepts relative to node X_5.

of X_5; X_7 and X_8 are the children of X_5; X_7, X_8, X_9, X_{10}, and X_{11} are the descendants of X_5; X_6 is neither an ancestor nor a descendant of X_5; the sequence of nodes (X_5, X_4, X_2, X_6) forms an undirected path between X_5 and X_6; X_1, X_4, X_5, and X_3 form an undirected loop.

Each node X_i in a Bayesian network is associated with a conditional probability table (CPT) defining the distribution of that node given its parent set Π_i, $P(X_i|\Pi_i)$ (Heckerman, 1999). The numbers defining these conditional distributions are also referred to as the BN parameters. The joint distribution of a BN is defined as a product of factors as follows:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\Pi_i) \ . \tag{2.1}$$

This factorized definition of the probability distribution entails the *local Markov property*, that X_i is conditionally independent of its non-descendants given its parents Π_i. This property is a special case of *d-separation*, concept that is used to determine conditional independences from observing the BN structure alone. For two variables X_a and X_b, and a set of variables $\mathbf{W}$, $(X_a \perp X_b|\mathbf{W})$ holds whenever X_a and X_b are d-separated by $\mathbf{W}$. Whether X_a

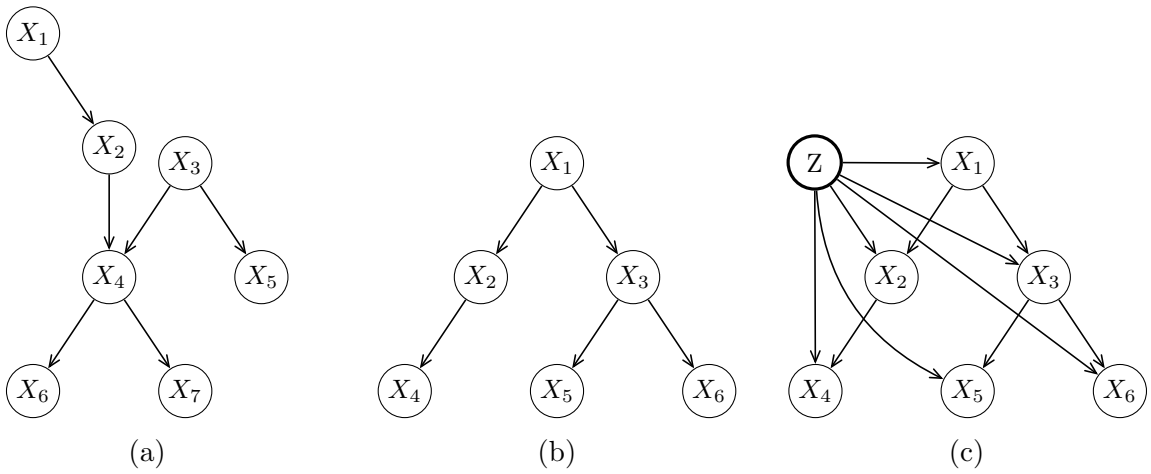

Figure 2: Examples of various special BN structures: (a) a polytree, (b) a Chow-Liu tree, and (c) a TAN.

and X_b are d-separated by $\mathbf{W}$ is determined as follows. Let a trail through the BN structure be defined as a (potentially self-intersecting) path over the undirected structure of the BN. A trail from a node X_a to a node X_b in the graph d-connects the nodes with respect to a variable set $\mathbf{W} \subset \mathbf{X}$ if for any triple of nodes $X_\alpha, X_\beta, X_\gamma$ in the trail, either $X_\beta \in \mathbf{W}$ and the arc directions are $X_\alpha \rightarrow X_\beta \leftarrow X_\gamma$, or $X_\beta \notin \mathbf{W}$ and the arc directions are in any other configuration. X_a and X_b are d-separated by $\mathbf{W}$ if there exists no trail that connects X_a to X_b (Geiger et al., 1990b).

There are several classes of BN structures that are notable to consider. Figure 2a shows an example of a *polytree*, also called a singly-connected network (SCN). A network is a polytree iff there exists no more than one undirected path between any pair of nodes in the network. An important property of polytree BNs is that they allow exact inference in polynomial time. Figure 2b shows an example of a *Chow-Liu tree*. A Chow-Liu tree is a polytree in which every node has at most one parent, thus forming a tree (or multiple trees) with arcs directed away from the root node. Chow and Liu (1968) showed that Chow-Liu trees can be optimally learned in polynomial time. Figure 2c shows an example of the

tree-augmented naive Bayes networks structure. The tree-augmented Naive Bayes (TAN) network structure is often used for classification. There is a class variable Z with variables as children. Those children form a tree structure. Friedman et al. (1997) showed that this structure can be learned efficiently.

2.3 BAYESIAN DIRICHLET FRAMEWORK FOR BAYESIAN NETWORKS

A Bayesian approach for estimating the probability distributions from the data is often applied in the context of BNs. This is accomplished by treating BN parameters as random variables with a posterior distribution rather than as point estimates. The idea of considering the parameters of a BN to be governed by a prior Dirichlet distribution has been used in multiple previous works (Heckerman et al., 1995), especially in the context of Bayesian scoring metrics for BN structures. Formally, this can be represented by the notation

$$P(X_i = x_{ik} | \Pi_i = \pi_{ij}) \equiv \Theta_{ijk} \ . \tag{2.2}$$

In the context of a Bayesian Dirichlet measure, such as K2 (Cooper and Herskovits, 1992) or BDeu (Heckerman et al., 1995), the prior distribution of $\boldsymbol{\Theta}_{ij} = (\Theta_{ij1}, \ldots, \Theta_{ijK_i})$ is Dirichlet (Johnson et al., 2002), that is, the probability density of $\boldsymbol{\Theta}_{ij}$ is

$$p(\boldsymbol{\theta}_{ij}) = \frac{\Gamma\left(\sum_{k=1}^{K_i} \alpha_{ijk}\right)}{\prod_{k=1}^{K_i} \Gamma\left(\alpha_{ijk}\right)} \prod_{k=1}^{K_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{2.3}$$

where $\Gamma(\cdot)$ is the Gamma function and α_{ijk} are the parameters of the Dirichlet distribution. This is also denoted by the notation

$$\boldsymbol{\Theta}_{ij} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{ij}) \ . \tag{2.4}$$

Since the Dirichlet is a conjugate of the multinomial, the posterior distribution of the parameters given N points of data $\mathcal{D}$ is also Dirichlet, specifically:

$$\boldsymbol{\Theta}_{ij} | \mathcal{D} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{ij} + \mathbf{N}_{ij}) \tag{2.5}$$

where $\mathbf{N}_{ij} = (N_{ij1}, \ldots, N_{ijK_i})$ and N_{ijk} is the number of points for which $X_i = x_{ik}$ and $\Pi_i = \pi_{ij}$ in the data. Because of the convenient updating pattern where each Dirichlet parameter α_{ijk} in the prior distribution is updated by the corresponding count N_{ijk} to arrive at the corresponding posterior distribution parameter, α_{ijk} are also sometimes called "pseudocounts."

This dissertation focuses on the specific case where the prior $\mathbf{\Theta}_{ij}$ distributions are uniform, that is, where $\alpha_{ijk} = 1$ for all values of i, j, k. In spite of this, all the methods used are readily generalizable to other prior Dirichlet distributions.

2.4 DATA USED FOR CASE STUDIES IN THIS DISSERTATION

Throughout the dissertation I present case studies to illustrate the explanation methods I develop. The following is an overview of the data used.

2.4.1 CEHC-PORT

The data used were collected in a prospective cohort study of hospitalized and ambulatory care patients conducted from October 1991 to March 1994 at five medical institutions (Kapoor, 1996). Patients included in the study had to have one or more symptoms suggestive of pneumonia, as well as radiographic evidence of pneumonia within 24 hours of presentation. The variables available in the data include categorical variables, continuous variables, and discretized versions of continuous variables. We restricted ourselves only to categorical variables and one discretization of each continuous variable, yielding 165 variables. The available variables included demographic information, history and physical examination information, laboratory results, chest X-ray findings, and outcomes. In using this data I selected two of the five medical institutions that participated in the study as the two data groups to compare.

3.0 RELATED WORK

This chapter is an overview of related work. While, to my knowledge, this approach to explaining distributional differences, especially in terms of contrast points, is novel, there are many works that are relevant to the approach, including BN learning (Section 3.1) and explanation (Section 3.3), as well as other statistical methods (Section 3.4).

3.1 BN LEARNING

A central component to the approach in this dissertation is the construction of a model of the data. Learning the structure and parameter distribution of a BN model from data is a necessary step in the difference detection methods I discuss in Chapter 4. In some situations the structure or elements of the structure may be known, but in many situations the structure must be learned from the data. This section discusses relevant previous work BN learning. Learning a BN is an important step for detecting group differences and producing the model used in the explanation, but it is merely a step, and the analysis methodology that I develop in this dissertation work goes beyond that step. Nevertheless, the choice of BN structure has a significant effect on guiding the resultant analysis.

Many approaches to BN learning have been explored in the literature. Daly et al. (2011) provide an extensive review of BN learning and divide existing methods across multiple categories: score search, where the space of BN structures is searched for a structure that has the best score according to some scoring criterion; constraint-based methods, where conditional independencies (CI) in the data are used to constrain the structure; and dynamic programming, which would have been more appropriately named "exact score optimization"

as opposed to the other score search methods, which are heuristic and do not guarantee optimal solutions.

Constraint-based methods use CIs obtained from statistical tests on the data to eliminate possible arcs in the network structure. Some of the earliest work is by Geiger et al. (1990a), who developed an algorithm to recover polytrees from an oracle which can determine if two variables are conditionally independent from each other. Systems for recovering DAGs from data came later, with the SGS algorithm by Glymour et al. (1991) and its variations, such as the PC algorithm by Spirtes and Glymour (1991). More recent work by Kalisch and Bühlmann (2007) shows the applicability of PC to high-dimensional data. Related algorithms for learning DAGs were also developed by Verma and Pearl (1992), and following the large body of work on finding DAGs using CIs, many refinements for specialized situations were developed by other authors (Bell et al., 2002; Cheng et al., 1997; Cooper, 1997; de Campos, 1998; de Campos and Huete, 1997, 2000b; Gou et al., 2007; Margaritis and Thrun; Meek, 1995).

Although these CI methods are appealing and often preferred for building causal BNs because of the explicit CI tests used for finding relationships, they may require large sample sizes to learn BNs effectively, and typically perform better when the graph to be found is sparse, that is, when there are many CIs in the data. Conversely, score search methods typically require less data to perform well and perform better on data that admit dense graphs (Daly et al., 2011).

Score search techniques seek to optimize some score function of the graph based on the data. The space over which most methods in this category search is that of possible DAGs, however, there are some methods that search over variable orderings (Acid and De Campos, 1996; Chen et al., 2008; de Campos and Huete, 2000a; De Campos et al., 2008; Larrañaga et al., 1996; Teyssier and Koller, 2012; Wallace and Korb, 1997), and some that search over equivalence classes[1] of BNs (Chickering, 2002; Chickering and Meek, 2006; Nielsen et al., 2002).

The space of DAGs is combinatorial in the number of variables, which makes it infeasible

[1] It is often possible to represent the same set of CIs using multiple different BN structures. Every such collection forms an *equivalence class* of BNs that corresponds to a particular set of CIs.

to exhaustively search it for the optimally-scoring structure in most cases. In fact, Chickering (1996) proves that such optimization is NP-hard in the general case. Hence, most search methods apply various heuristics and do not perform an exhaustive search of the space, and cannot guarantee to find the optimal structure. Notable exceptions, which are feasible in the case of up to approximately 30 variables, include dynamic programming approaches (Koivisto and Sood, 2004; Silander and Myllymaki, 2006; Singh and Moore, 2005) and an application of A* search to the space of DAGs by Yuan et al. (2011).

Various heuristic search methods have been applied for learning BNs. All methods operate on the common framework of applying operators to move from one DAG to another, such as arc additions, removals, and reversals. Using these operators, the space of DAGs is traversed until some criterion is met and the "best" structure is selected. The K2 algorithm described by Cooper and Herskovits (1992) is one of the earliest works to use a greedy search algorithm. It used the Bayesian scoring criterion that came to be known as the K2 score, and required an ordering over the variables. Other greedy search approaches have been developed, such as K3 by Bouckaert (1993), and others more recently (Liu and Zhu, 2007; Liu et al., 2007). Greedy methods that do not require an ordering over the variables have also been developed (Chickering and Meek, 2006; de Campos et al., 2002; Hwang et al., 2002). In this dissertation I use greedy-thick-thinning—a similar method that maximizes the K2 score in a greedy fashion by starting with an empty graph, adding arcs that most increase the score until no more arc additions can increase the score, and then performs arc removals that increase the score most until no more removals increase the score. Another notable approach to search that has been applied is genetic algorithms (GA). Larrañaga et al. (1996) used GAs to search over orderings and used K2 to find a DAG for each ordering. Others, such as Wong et al. (1999) applied GAs directly to the DAG space. Hybrid approaches that combine evolutionary algorithms with other techniques have also be explored, such as (Wong and Leung, 2004).

All of the score search methods attempt to optimize some particular score of the DAG structure that is based on data and prior belief. Various scores that have been used in the literature include the Bayesian Dirichlet (BD) criterion, the Akaike information criterion (AIC), and the Bayesian information criterion/minimum description length (BIC/MDL).

The main feature that all of these measures have in common is that they reward simpler structures and a better fit to the data.

This dissertation uses BD scoring because of the natural interpretation of the score as a posterior probability, and because it enables the treatment of network parameters as random variables. Using these scores essentially consists of choosing the graph structure $\mathcal{M}$ that is most probable given the data $\mathcal{D}$ and prior belief. By applying Bayes' rule we obtain that

$$P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{M}, \mathcal{D}) = P(\mathcal{M})P(\mathcal{D}|\mathcal{M}) \tag{3.1}$$

where $P(\mathcal{M})$ is a prior for the graph structure. Often the graph structure prior is assumed to be uniform, and the goal becomes to maximize $P(\mathcal{D}|\mathcal{M})$, the probability of the data given the structure. Under the Bayesian Dirichlet framework outlined in section 2.3, this is computed in closed-form as

$$\begin{aligned} P(\mathcal{D}|\mathcal{M}) &= E_{\boldsymbol{\Theta}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}, \mathcal{M}) = \\ &= \int_{\boldsymbol{\theta}} p(\mathcal{D}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot} + N_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad . \end{aligned} \tag{3.2}$$

Different choices of the Dirichlet priors lead to different BD scores: for example, the K2 score (Cooper and Herskovits, 1992) is obtained from using uniform priors (all $\alpha_{ijk} = 1$), and the BDeu score Heckerman et al. (1995) is obtained from using priors with $\alpha_{ijk} = \frac{\alpha^*}{J_i K_i}$ where α^* is a hyperparameter known as the Equivalent Sample Size (ESS).

The methods developed in this dissertation are presented at a level of generality compatible with any Bayesian Dirichlet scoring measure and any structure search method that aims to maximize the measure. This dissertation applies the heuristic greedy-thick-thinning search method for learning the BN structure using the K2 score in all evaluations and applications of methods developed to data.

3.2 WHAT IS "EXPLANATION"

A considerable body of literature on explanation exists in the context of artificial intelligence and expert system development. Explanation is essential to productive communication between a human user and an intelligent system in order for the user to be able to understand and trust the system's conclusions. I believe that user understanding is equally important when the system is not prescriptive or predictive (as intelligent and expert systems usually are), but rather, descriptive, as in this dissertation. Lacave and Diez (2000) conclude that explanation consists in "exposing something in such a way that it is understandable for the receiver of the explanation – so that he/she improves his/her knowledge about the object of the explanation – and satisfactory in that it meets the receiver's expectations." The "exposition" component of explanation clearly maps to the tasks of parameter difference detection and of showing how those differences interact. The goal of making this exposition "satisfactory" has a considerable impact on the considerations of how explanations are generated in the methods of Chapter 6 and the design considerations of the prototype in Chapter 7.

3.3 BN EXPLANATION

Most explanation methods are designed around the task of explaining inference. The task of BN inference is that of computing a posterior probability $P(\mathbf{d}|\mathbf{e})$, where $\mathbf{e}$ is evidence – an assignment of observed values to a subset of the variables in the network – and $\mathbf{d}$ is the assignment of one or more of the other variables in the network, typically representing the prediction of an event of interest, such as a disease diagnosis. Two major approaches to inference explanation are abduction and influence-tracing. Abduction provides an assignment of unobserved variables to values, usually the most probable values, essentially providing a plausible scenario as an explanation. These methods include methods of total abduction (Charniak and Shimony, 1994; Gámez, 2004; Pearl, 1988), where all unobserved variables are assigned values, and methods of partial abduction (Gámez, 2004; Shimony, 1991), where only variables that are relevant to the particular inference task are assigned values.

Influence-tracing methods describe the relationships between variables in terms such as positive or negative associations between variables and the strengths of those associations (Lacave and Diez, 2000). There are influence-tracing methods for describing influences qualitatively (Druzdzel, 1993), which use verbal descriptions or visual cues to indicate whether the relationship between one variable and another is positive or negative (whether higher values of one variable correlate with higher values of the other), whether the relationships are strong or weak, and whether a pair of variables is conditionally independent given the state of another variable. There are also influence-tracing methods that quantify how a change in the state of one variable affects the inference about the probable state of another. Measures used to quantify such changes include differences of probabilities (Lacave et al., 2007), log-ratios of probabilities (Madigan et al., 1997), and other functions such as cross-entropy (Suermondt and Cooper, 1993; Suermondt, 1992). The difference explanation methods in Chapter 6 are similar to influence-tracing methods in that they also quantitatively compare probabilities that are obtained by conditioning on the group compared. This approach is novel since it aims at a descriptive analysis of the distribution of the data, which is considerably different from the task of explaining the inference behind computing a posterior probability. Moreover, the methods here provide an organized explanation of relevant differences in terms of the underlying parameter differences, which is a task that has not appeared in previous explanation work.

3.4 OTHER STATISTICAL METHODS

There are various statistical methods that are applicable to the problem of identifying differences across a pair of groups. The statistical approach that most closely relates to the difference detection task set forth in this dissertation is that of *contrast set mining*. Bay and Pazzani (2001) present contrast-set mining as the discovery of joint variable-value assignments that have different levels of support in different groups. The approach taken parallels association-rule mining in that the space of possible joint variable-value assignments is searched to maximize a score (in association-rule mining, this score is the lift of a rule, while

in contrast-set mining a chi-square test is used). The main challenge in contrast-set mining is the search of the exponentially large space of possible sets (joint variable-value assignments), and much of the literature is dedicated to discussing heuristics and pruning rules to make the search feasible. The output of contrast-set mining is the list of joint variable-value assignments (the sets) which have differing support across groups, ranked by the extent of that difference and tested for significance. Webb et al. (2003) show that contrast-set mining can be viewed as a special case of general association-rule mining, and later work by Novak et al. (2009) relates the task to emerging pattern mining and subgroup mining. Somewhat analogously, in Section 4.2 I show that under certain assumptions, difference detection is a special case of BN structure learning. In a sense, even the less restrictive approach of Section 4.3 can be viewed as a special case learning, specifically the learning of a BN structure with context-specific independence, as discussed in Section 8.2.7.

It is important to note that while contrast-set mining is well-suited for characterizing the differences across two groups of data, since it does not build a model of the data, there is no framework for the explanation of how various differences relate to each other using contrast sets.

Another natural statistical approach to difference detection is to treat the problem as a classification problem, that is, let Z be a group indicator variable (as in 4.2) and find a good set of predictors for Z. The reasoning is that the predictors of Z are those variables that "behave differently" across the two groups. There are many different classification methods that can perform such a task. Rule learning and frequent pattern mining provide logical decision rules that describe which value of Z a record is most likely to match. In a similar way, classification and regression trees (CART) can, for a given record, based on whether the values of the predictors fall into specific ranges, say which value of Z it matches. Perhaps the most widely used classification method is logistic regression, which builds a linear model of the log-odds of a record corresponding to a particular value of Z. The linearity of the model makes it possible to explain the model in terms of additive contributions of each prediction to the log-odds (Poulin et al., 2006). The application of logistic regression for this task has become fairly standard practice, for example Cleophas and Zwinderman (2012) recommend using it as a means for post-hoc analysis in clinical studies. Poulin et al. (2006) also show

that a similar explanation can be produced for other additive classifiers, such as support vector machines and Naive Bayes models.

The main issue with approaching the problem as classification is that it answers a different question from the one posed: the classification task is to find a small set of variables that separates the two groups well, while the difference explanation task is to identify all those variables that are significantly different between the groups. While these tasks are closely related it is important to recognize that they are not the same task. Most classification methods attempt to predict the class variable, the group, in this analogy, using a minimal number of predictors, that is, if one predictor explains the difference there is no need to introduce another predictor. In this sense, classifiers are good at identifying relationships between predictors and group membership, modeling problems may arise when predictors are highly correlated, or alternatively, all but one of a collection of correlated predictors may be dropped. Hence, when the task, as it is here, is actually to find similarities and differences among the relationships between the predictors themselves, the classification approach is limited.

There is one type of classifier that *can* be applied almost directly to the task of difference explanation. Various augmented Naive Bayes classifiers such as TAN (Friedman et al., 1997). What is special about those models is that they build a model for the distribution of all predictors given the classification variable. We can then use the learning phase of such a classifier to learn the BN structure that defines distributions of the predictors given the group variable, thereby giving us the parameter differences in each predictor variable across the two groups. The main limitation of using augmented Naive Bayes classifiers in this manner is that since they are optimized for prediction, the types of relationships they find between other variables in the data are limited to those necessary to improve that prediction task. This is in contrast to our focus on relationships between variables in the data in the task of explaining differences across a pair of groups.

Traditional statistical methods can also be applied to detect differences in the distributions of two groups. There are multiple traditional statistical tests that are designed to compare distributions. For categorical variables, the Chi-Square test is applicable, it tests whether two groups are independent. This can be used to determine if a variable has differ-

ent distributions across two groups by testing whether it is dependent on the group variable. More generally, for continuous variables, the Kolmogorov-Smirnov test is often used to determine equality of distributions. Note that these tests are univariate, and cannot therefore be used to compare two multivariate groups of data directly. There are other measures of distribution differences that are multivariate in nature, such as Hotelling's T-squared test, mutual information, or Kullback-Leibler divergence (Kullback and Leibler, 1951). These measures are multivariate, but they do not allow for examining the contributions of differences in individual variables to the overall measure of difference across the groups. The approach in this dissertation bridges this gap with a method that can be used as a measure of overall difference, while allowing for the contributions in the differences of individual variables to the overall measure of difference to be quantified.

4.0 DETECTING DIFFERENCES IN DISTRIBUTIONS

This dissertation approaches the problem of detecting differences from a statistical standpoint, where given a pair of data groups $\mathcal{D}_1$ and $\mathcal{D}_2$ over a vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$, we would like to identify variables that exhibit statistical differences. A variable might have a different marginal distribution in the two groups and/or a different conditional distribution when conditioning on the values of some of the other variables.

To clearly and simply illustrate goals and concepts of difference detection, consider the "Balloons" data set from the UCI Machine Learning Repository (Bache and Lichman, 2013). The data were originally used in an experiment about knowledge acquisition where subjects were shown photographs of a person doing something with a balloon. The balloon in the photograph might be yellow or purple, large or small; the person in the photograph might be an adult or a child, and they might be dipping or stretching the balloon, and either inflating or not inflating the balloon, based on some rule. In the knowledge acquisition experiment, the task of the subjects was to predict whether the person in the photograph would inflate the balloon (Pazzani, 1991). The data represents information from the photographs shown to the subjects in a pair of experiments, and consist of the five binary variables *color, size, act, age*, and *inflated*, which take the value pairs *yellow/purple, large/small, stretch/dip, adult/child*, and *T/F*, respectively. We will focus on two groups of data labeled 'adult-stretch' and 'adult+stretch' in the UCI data repository. I will refer to the two groups as the **or** group and the **and** group, respectively. The difference between the two data groups is that in the **or** group, *inflated* is *T* iff *age* is *adult* **or** *act* is *stretch*, while in the **and** group, *inflated* is *T* iff *age* is *adult* **and** *act* is *stretch*. When comparing these two groups, conceptually, this is the relationship that we wish to extract. In a statistical sense, we want to identify that *inflated* is the variable that is behaving differently between the two groups,

and moreover, that it is the conditional distributions of *inflated* given *age* and *act* that are different between the two groups.

4.1 A CONCEPTUAL FRAMEWORK

Let us take the task illustrated with the Balloons example above and generalize it to an empirically testable hypothesis.

4.1.1 Difference recovery hypothesis

Let us view $\mathcal{D}_1$ and $\mathcal{D}_2$ as two groups of data generated from two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The models might be statistical or (as in the Balloons example) rule-based. The models may differ in various ways, for example, the value of a variable X_i may depend on one set of variables in $\mathcal{M}_1$ and a different set of variables in $\mathcal{M}_2$. Another type of difference that is of interest happens when the value of X_i depends on the same set of variables in both models, but, if the model is rule-based, the rule for determining X_i is different between the models, or if the model is statistical, the distribution of X_i is different between the models. I hypothesize that such differences are reflected in the data $\mathcal{D}_1$ and $\mathcal{D}_2$, and that the differences between the models can be recovered from the data. The statistical difference detection methods developed in this chapter and tested in the next test this hypothesis by demonstrating such recovery.

The difference recovery hypothesis is: *given $\mathcal{D}_1$ and $\mathcal{D}_2$, groups of data generated from models (either statistical or rule-based) $\mathcal{M}_1$ and $\mathcal{M}_2$, where there may be both differences and similarities between the models. I hypothesize that we can accurately recover a statistical representation of the differences and similarities between the models from the data $\mathcal{D}_1$ and $\mathcal{D}_2$.*

4.1.2 Comparing parameters across models

Consider two groups of data $\mathcal{D}_1$ and $\mathcal{D}_2$ over the same set of variables $\mathbf{X} = (X_1, \ldots, X_n)$, and denote the concatenation of $\mathcal{D}_1$ and $\mathcal{D}_2$ by $\mathcal{D}_\cup$. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are not different in a statistical sense, they follow the same distribution, which is therefore the distribution of $\mathcal{D}_\cup$. Let $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_\cup$ denote maximum *a posteriori* (MAP) models within some space of models for the data in $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_\cup$ respectively. A MAP model $\mathcal{M}$ given data $\mathcal{D}$ is a model that maximizes the likelihood $P(\mathcal{D}|\mathcal{M})$. In the case where $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same, we expect that $P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2) \leq P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ in the large sample limit, since modeling the two groups as governed by independent distributions does not yield a better fitting model than when the groups are modeled as coming from the same distribution. In the case where $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically different, we expect $P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2) > P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ in the large sample limit.

To address the problem posed by the difference recovery hypothesis of Section 4.1.1, I extend this idea from the model level to the parameters of the models, an extension that can be applied when the models have the following properties: the distribution of a variable X_i is defined by a vector of parameters $\boldsymbol{\theta}_i$, parameters $\boldsymbol{\theta}_i$ are drawn from a distribution $\boldsymbol{\Theta}_i$, and parameter independence holds, such that, $\boldsymbol{\Theta}_i \perp \boldsymbol{\Theta}_{i'}$ for $i \neq i'$. BNs with Dirichlet parameter priors have these two properties. These properties provide a means to match parameters from $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ and pose the question of whether a variable X_i is distributed differently across the two groups as that of whether

$$E_{\boldsymbol{\Theta}_i|\mathcal{M}_1}[P(\mathcal{D}_1|\boldsymbol{\Theta}_i, \mathcal{M}_1)] \times E_{\boldsymbol{\Theta}_i|\mathcal{M}_2}[P(\mathcal{D}_2|\boldsymbol{\Theta}_i, \mathcal{M}_2)] > E_{\boldsymbol{\Theta}_i|\mathcal{M}_\cup}[P(\mathcal{D}_\cup|\boldsymbol{\Theta}_i, \mathcal{M}_\cup)] \ . \tag{4.1}$$

In the following two sections, I explore two approaches to applying this general framework to BNs with Dirichlet priors, one that explicitly uses only one BN model, and is conceptually equivalent to enforcing a shared structure for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$, and one that is less restrictive, and only requires that the three models share a variable ordering.

4.2 UNI-MODEL APPROACH

This section introduces an approach that provides a direct matching between the parameters of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ by ensuring they share a single structure. First I show that the task of constructing this trio of models becomes equivalent to learning a single BN model with group indicator variable Z that takes one value for records from $\mathcal{D}_1$ and another for records from $\mathcal{D}_2$, subject to the constraint that Z is an orphan in the network structure. Next, I discuss the implications of this orphan constraint for interpreting the learned BN. Finally, I discuss the application of this model to the detection of statistically and clinically significant differences.

4.2.1 Equivalence to learning a single BN model

Recall that generally, in a BN, the distribution of a variable X_i is defined in terms of a set of conditional distributions $X_i|\pi_{ij}$, one for each variable assignment π_{ij} of the parents Π_i of X_i. Thus, given data $\mathcal{D}$, if we have a model structure $\mathcal{M}$ that specifies the parent set Π_i, we can obtain the posterior distribution the parameter vector $\mathbf{\Theta}_i$ that defines the conditional distribution of X_i, and consequently we can compute the quantity of interest

$$E_{\mathbf{\Theta}_i|\mathcal{M}}[P(\mathcal{D}|\mathbf{\Theta}_i,\mathcal{M})] = \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot}+N_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \quad . \tag{4.2}$$

Coming back to the problem of testing inequality (4.1) subject to the constraint of identical structure, (4.2) gives us a direct way to compute each term. Particularly, note that, denoting counts in $\mathcal{D}_1$ and $\mathcal{D}_2$ by $N^{(1)}$ and $N^{(2)}$, respectively, we obtain:

$$\begin{aligned} &E_{\mathbf{\Theta}_i|\mathcal{M}_1}[P(\mathcal{D}_1|\mathbf{\Theta}_i,\mathcal{M}_1)] \times E_{\mathbf{\Theta}_i|\mathcal{M}_2}[P(\mathcal{D}_2|\mathbf{\Theta}_i,\mathcal{M}_2] = \\ &\quad = \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot}+N^{(1)}_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk}+N^{(1)}_{ijk})}{\Gamma(\alpha_{ijk})} \times \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot}+N^{(2)}_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk}+N^{(2)}_{ijk})}{\Gamma(\alpha_{ijk})} \quad . \end{aligned} \tag{4.3}$$

Note that for $z \in \{1,2\}$, $N^{(z)}_{ijk}$ is defined as the number of records in $\mathcal{D}_z$ for which $X_i = x_{ik}$ and $\Pi_i = \pi_{ij}$. If we add a group indicator variable Z to the model, as described above,

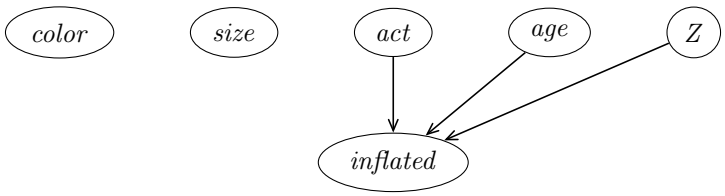

Figure 3: The uni-model BN learned for the Balloons example.

letting $\Pi_i' = \Pi_i \cup \{Z\}$ and denoting the parameters defining X_i given this new parent set by $\mathbf{\Theta}_i'$, we get that

$$E_{\mathbf{\Theta}_i|\mathcal{M}_1}[P(\mathcal{D}_1|\mathbf{\Theta}_i,\mathcal{M}_1)] \times E_{\mathbf{\Theta}_i|\mathcal{M}_2}[P(\mathcal{D}_2|\mathbf{\Theta}_i,\mathcal{M}_2] =$$
$$= \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot}+N_{ij\cdot}^{(1)})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk}^{(1)})}{\Gamma(\alpha_{ijk})} \times \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot}+N_{ij\cdot}^{(2)})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk}^{(2)})}{\Gamma(\alpha_{ijk})} =$$
$$= \prod_{j'=1}^{J_i'} \frac{\Gamma(\alpha_{ij'\cdot})}{\Gamma(\alpha_{ij'\cdot}+N_{ij'\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ij'k}+N_{ij'k})}{\Gamma(\alpha_{ij'k})} = E_{\mathbf{\Theta}_i'}[P(\mathcal{D}_\cup|\mathbf{\Theta}_i')] \ . \quad (4.4)$$

This means that when we use a score based on Dirichlet priors to learn a BN structure, the learning algorithm performs our desired comparison between $(\mathcal{M}_1, \mathcal{M}_2)$ and $\mathcal{M}_\cup$ on a local level whenever Z is considered as a candidate parent for a variable. We can ensure that Z is always considered as a candidate by constraining it to be an orphan in the graph. With this constraint, Z can potentially be a parent for any variable without violating the acyclicity constraint of the BN structure.

Consider applying this to the Balloons example, with Z indicating whether the data comes from the **or** group or the **and** group. K2 structure learning yields the BN in Figure 3. The *inflated* variable is correctly characterized as dependent on *act*, *age* and Z. It is also the only variable dependent on Z. This means that *inflated* is correctly identified as the variable the distribution of which is different across these two groups.

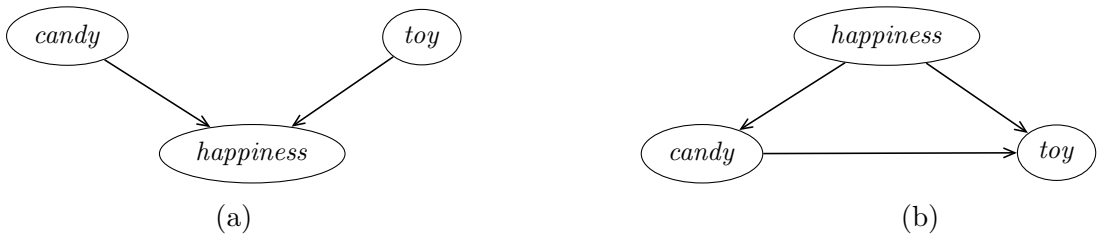

Figure 4: (a) Causal and (b) orphan-constraint-enforced graph structures for the candy-toy-happiness example.

4.2.2 The orphan constraint on Z

This section discusses the appropriateness and implications of imposing the constraint that Z must be an orphan. This constraint is justified when treating Z as a mathematical tool to help identify parametric differences between two BNs of identical structure. However, since Z is an added variable, it is important to understand the implications of including it as an orphaned variable among the domain variables. There are a few situations where it is causally appropriate for the variable Z to be orphaned, e.g. when it represents one time period as opposed to another, or when it is the treatment in a randomized controlled trial. In such situations, we are justified in assuming that no other variables can influence the value of Z. Clearly, we do not want to be limited only to those cases, and I claim that imposing the orphan constraint is indeed appropriate outside of those cases, provided that we interpret the model properly.

In order to understand why the constraint is appropriate, it is important to understand the apparent problem that is introduced by deviation from a causal graph. Suppose, for example, that our data are about children and that we have three variables: *toy*, *candy*, and *happiness*, indicating whether a child has received a toy, has received candy, or is happy, respectively. Furthermore, suppose that children receive toys and candy independently, and that both toys and candy have positive effects on happiness. The causal BN here would have *happiness* as a child of both *toy* and *candy*, with no arc between *toy* and *candy*, as shown in

Figure 4a. If we are tasked with the project of comparing the happy children in the data to the unhappy ones, the group division dictates that $Z = happiness$, and that the *happiness* node must be orphaned. The likely network that would be learned under that constraint is one like that in Figure 4b, with both *toy* and *candy* as children of *happiness* as well as an additional arc between *toy* and *candy*. This additional arc is representative of a dependency that appeared: given that a child is happy, knowing that they received candy informs us that it is less likely that they also received a toy, since the candy already explains the happiness. The apparent problem with this learned structure is that it seems to contradict our initial statement that children receive candy and toys independently, suggesting that the relationship we found between *toy* and *candy* is not real.

It is important to interpret statistical relationships in the appropriate context. In the context of a known *happiness* state, there is a very real statistical relationship between *toy* and *candy* caused by conditioning on the *happiness* state. There is a dependence relationship between *toy* and *candy* in the set of happy children, and there is a dependence relationship between *toy* and *candy* in the set of unhappy children, and it is important to include that relationship in the model. Similarly, relationships found when learning a BN with an orphan constraint on the group indicator Z in any data may be ones due to confounding by Z. These relationships exist within each of the two groups that are compared, even if these are relationships between variables that are causally independent in the broader scope of the entire data taken jointly.

Section 8.2.1 discusses some ideas for augmenting a network learned under such constraints with an explicit differentiation between relationships due to statistical dependence and causation.

4.2.3 Statistical and clinical significance for differences

When data are assumed to follow some random distribution, some variability is expected. Even if two groups of data are generated from the same distribution, we would almost always expect to see some differences in the data. There are two notions of significance that are often used to evaluate an effect (in this case, a difference) in the data: statistical significance

and clinical significance.

An effect is considered statistically significant if there is sufficient evidence to conclude that the effect is not due to random variability Coolidge (2012). An effect is considered clinically significant if the effect size is large enough to matter in the practical sense Kazdin (1999).

Note that the process of learning a BN structure, as described above, already performs a sort of Bayesian statistical significance test. As discussed, when the BN structure learning algorithm considers whether to add Z to the set of parents of X_i, it compares the likelihood of the data subject to a model that allows different parameters for the two groups to a likelihood of the data subject to a model that uses the same parameters for the two groups. This constitutes a Bayesian statistical significance test, the outcome of which is reflected in whether Z is a parent of X_i in the final learned BN.

Consider now a few approaches to testing whether a difference between parameters is clinically significant. The most basic test is to examine whether a difference or a ratio between the parameter estimates exceeds some threshold:

$$|\theta_{ijk}^{(2)} - \theta_{ijk}^{(1)}| > \epsilon \ \text{, or} \tag{4.5}$$

$$|\log \theta_{ijk}^{(2)} - \log \theta_{ijk}^{(1)}| > \epsilon \ . \tag{4.6}$$

Within a probabilistic framework, we can consider the posterior distribution of the parameters, and compute the probability of the difference or ratio exceeding that threshold:

$$P(|\Theta_{ijk}^{(2)} - \Theta_{ijk}^{(1)}| > \epsilon) > 1 - \alpha \ \text{, or} \tag{4.7}$$

$$P(|\log \Theta_{ijk}^{(2)} - \log \Theta_{ijk}^{(1)}| > \epsilon) > 1 - \alpha \ . \tag{4.8}$$

This probabilistic test combines statistical and clinical significance by evaluating not only whether the effect size is sufficient, but also whether the observation that the effect size is sufficient is itself not due to random variability.

I report the performance of these tests on the task of difference detection in Chapter 5.

4.3 MULTI-MODEL APPROACH

The uni-model approach constrained the models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ to share the same model structure. The multi-model approach relaxes the identical structure constraint to a variable ordering constraint. The ordering constraint can be enforced by learning $\mathcal{M}_\cup$ without constraints and then constraining $\mathcal{M}_1$ and $\mathcal{M}_2$ to have the same topological ordering of variables as does $\mathcal{M}_\cup$. There are many other possible approaches to enforcing these constraints, ranging from obtaining an order *a priori* to more sophisticated approaches that seek an order optimal for all three networks jointly. Exploring all of these alternatives is outside the scope of this work.

This section describes the resulting approach which takes the three learned structures, constrained as above, exploits parameter independence properties to match parameters, and derives a Bayesian score for detecting whether groups of data are different at the full-model level, at the variable level, and at a sub-variable (groups of parameters) level.

4.3.1 Measuring differences using Bayes factors

Given Dirichlet parameter priors over network parameters and uniform priors over network structures, the MAP models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ correspond to the structures learned by maximizing the BD scores for the data groups $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_\cup$ respectively, and, moreover, the probabilities $P(\mathcal{D}_1|\mathcal{M}_1)$, $P(\mathcal{D}_2|\mathcal{M}_2)$, and $P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ correspond to those scores. Hence, we can view the ratio

$$\frac{S}{T} = \frac{P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2)}{P(\mathcal{D}_\cup|\mathcal{M}_\cup)} = \frac{E_{(\boldsymbol{\Theta}^{(1)}|\mathcal{M}_1),(\boldsymbol{\Theta}^{(2)}|\mathcal{M}_2)} P(\mathcal{D}_1|\boldsymbol{\Theta}^{(1)}, \mathcal{M}_1) \times P(\mathcal{D}_2|\boldsymbol{\Theta}^{(2)}, \mathcal{M}_2)}{E_{\boldsymbol{\Theta}^{(\cup)}|\mathcal{M}_\cup} P(\mathcal{D}_\cup|\boldsymbol{\Theta}^{(\cup)}, \mathcal{M}_\cup)} \tag{4.9}$$

as a score that measures dissimilarity between the distributions governing $\mathcal{D}_1$ and $\mathcal{D}_2$. The numerator $S = P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2)$ denotes the likelihood of the data subject to the hypothesis that the two groups are modeled independently, and the denominator $T = P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ denotes the likelihood of the data subject to the hypothesis that the two groups are modeled as coming from the same distribution. Since each Bayesian score

$P(\mathcal{D}|\mathcal{M})$ is obtained by averaging over a parameter space, the ratio S/T is the Bayes factor (Jeffreys, 1998) for comparing these two hypotheses.

Another way to think of the hypotheses associated with the likelihoods S and T is in terms of whether the parameter set $\boldsymbol{\Theta}^{(1)}$ for group 1 is independent or whether it is identical to the parameter set $\boldsymbol{\Theta}^{(2)}$.

$$S = P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}^{(1)} \perp \boldsymbol{\Theta}^{(2)}) \quad \text{and} \tag{4.10}$$

$$T = P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)}) \ . \tag{4.11}$$

Due to global and local parameter independence, these likelihoods can be seen as a product of per-variable likelihoods, each of which is a product of likelihoods defined by even more local groups of parameters. This decomposition allows us to go further and measure contributions to this score on a per-variable basis, as well as further analyze the contributions of finer-scale local differences. Particularly:

$$E_{\boldsymbol{\Theta}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}, \mathcal{M}) = \prod_{i=1}^{n} E_{\boldsymbol{\Theta}_i|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_i, \mathcal{M}) \tag{4.12}$$

where

$$E_{\boldsymbol{\Theta}_i|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_i, \mathcal{M}) = \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot} + N_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \ . \tag{4.13}$$

This leads to the following decomposition of the Bayes factor (4.9):

$$\frac{S}{T} = \prod_{i=1}^{n} \frac{S_i}{T_i} \tag{4.14}$$

where

$$\begin{aligned} S_i &= P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}_i^{(1)} \perp \boldsymbol{\Theta}_i^{(2)}) = E_{(\boldsymbol{\Theta}_i^{(1)}|\mathcal{M}_1),(\boldsymbol{\Theta}_i^{(2)}|\mathcal{M}_2)} P(\mathcal{D}_1|\boldsymbol{\Theta}_i^{(1)}, \mathcal{M}_1) \times P(\mathcal{D}_2|\boldsymbol{\Theta}_i^{(2)}, \mathcal{M}_2) \\ T_i &= P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)}) = E_{\boldsymbol{\Theta}_i^{(\cup)}|\mathcal{M}_\cup} P(\mathcal{D}_\cup|\boldsymbol{\Theta}_i^{(\cup)}, \mathcal{M}_\cup) \ . \end{aligned} \tag{4.15}$$

The ratio S_i/T_i for a particular variable X_i is itself a Bayes factor that compares the two modeling hypotheses as they pertain to defining the probability distribution of X_i.

Consider the Balloons data set example. The network structures learned using greedy-thick-thinning with a K2 score turn out to be the same for $\mathcal{D}_1$ (16 cases), $\mathcal{D}_2$ (16 cases), and $\mathcal{D}_\cup$ (32 cases); that structure is shown in Figure 5. The value of S/T for this data is 4.873.

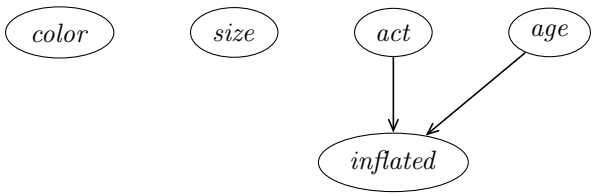

Figure 5: The network structure that the learned models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ share for the Balloons example.

Jeffreys (1998), and later Kass and Raftery (1995), provide guidelines for interpreting Bayes factors, and depending on the guideline used, this value indicates "substantial" (Jeffreys, 1998) or "positive" (Kass and Raftery, 1995) evidence in favor of modeling the distributions as different. The values of S_i/T_i are shown in Table 1.

We can see that the higher value of S_i/T_i (53.00) for the *inflated* variable correctly suggests that there is strong evidence in favor of the distribution of *inflated* being different across the two groups.

S_i/T_i can be further decomposed into terms that represent contributions to the likelihood from differences in the local graph structure of the MAP models and differences in the conditional counts in the data. Note that due to local parameter independence, the node-wise likelihood is a product of likelihoods over parent configurations for that node:

$$E_{\boldsymbol{\Theta}_i|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_i, \mathcal{M}) = \prod_{j=1}^{J_i} E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_{ij}, \mathcal{M}) \tag{4.16}$$

where

$$E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_{ij}, \mathcal{M}) = \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot} + N_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad . \tag{4.17}$$

In the case where the local structure coincides in the three models, that is, X_i has the same parents in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$, we can compute the ratio

$$\frac{E_{(\boldsymbol{\Theta}_i^{(1)}|\mathcal{M}_1),(\boldsymbol{\Theta}_{ij}^{(2)}|\mathcal{M}_2)} P(\mathcal{D}_1|\boldsymbol{\Theta}_{ij}^{(1)}, \mathcal{M}_1) \times P(\mathcal{D}_2|\boldsymbol{\Theta}_{ij}^{(2)}, \mathcal{M}_2)}{E_{\boldsymbol{\Theta}_{ij}^{(\cup)}|\mathcal{M}_\cup} P(\mathcal{D}_\cup|\boldsymbol{\Theta}_{ij}^{(\cup)}, \mathcal{M}_\cup)} \tag{4.18}$$

Table 1: Variable-wise scores obtained for the Balloons data.

Variable	Bayes factor S_i/T_i
inflated	53.00
act	0.8076
age	0.8076
color	0.3754
size	0.3754

to obtain a Bayes factor for whether $\mathbf{\Theta}_{ij}^{(\cdot)}$, the parameter set that defines $X_i|\pi_{ij}$, is different across the two groups. In the general case, however, the parent sets of X_i can be different in the three models, and may have some partial overlap. In that case, we base a decomposition of S_i/T_i on the possible configurations of the intersection of the parent sets of X_i in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$. Specifically, denote the parent sets of X_i in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ by $\Pi_i^{(1)}$, $\Pi_i^{(2)}$, $\Pi_i^{(\cup)}$ respectively. Let $\overset{\cap}{\Pi}_i$ denote $\Pi_i^{(1)} \cap \Pi_i^{(2)} \cap \Pi_i^{(\cup)}$. Let $J_i^{(\cdot)}$ be the number of possible configurations of $\Pi_i^{(\cdot)}$, and enumerate those configurations by $j = 1, \ldots, J_i^{(\cdot)}$. Let H_i be the number of possible configurations of $\overset{\cap}{\Pi}_i$ and and enumerate those configurations by $\eta = 1, \ldots, H_i$. For example, suppose that in data where all variables are binary, for a variable X_1 we have $\Pi_1^{(\cup)} = \{X_2, X_3, X_4\}$, $\Pi_1^{(1)} = \{X_2, X_3, X_5\}$, and $\Pi_1^{(2)} = \{X_2, X_4, X_5\}$. Then we have that $\overset{\cap}{\Pi}_1 = \{X_2\}$, and there are two possible configurations $\eta = 1$ and $\eta = 2$ for this set, corresponding to x_{21} and x_{22}. Let $J_i(\eta)$ indicate the subset of parent configurations $j \in \{1, \ldots, J_i^{(\cdot)}\}$ for the variables in $\Pi_i^{(\cdot)}$ that are consistent with configuration η. For example, if $\eta = 1$ represents x_{21} in our example, then $J_1^{\cup}(1)$ is the set of j-values that correspond to the set of parent assignments $\{(x_{21}, x_{31}, x_{41}), (x_{21}, x_{31}, x_{42}), (x_{21}, x_{32}, x_{41}), (x_{21}, x_{32}, x_{42})\}$.

Grouping product terms by those that correspond with individual values of η, yields

$$\frac{S_i}{T_i} = \prod_{\eta=1}^{H_i} \frac{S_{i\eta}}{T_{i\eta}} \tag{4.19}$$

with

$$
\begin{aligned}
S_{i\eta} =& P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} \perp \boldsymbol{\Theta}_{i\eta}^{(2)}) = \\
&= \left(\prod_{j \in J_i^1(\eta)} E_{\boldsymbol{\Theta}_{ij}^{(1)} | \mathcal{M}_1} P(\mathcal{D}_1 | \boldsymbol{\Theta}_{ij}^{(1)}, \mathcal{M}_1) \right) \left(\prod_{j \in J_i^2(\eta)} E_{\boldsymbol{\Theta}_{ij}^{(2)} | \mathcal{M}_2} P(\mathcal{D}_2 | \boldsymbol{\Theta}_{ij}^{(2)}, \mathcal{M}_2) \right) \\
T_{i\eta} =& P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)}) = \prod_{j \in J_i^{\cup}(\eta)} E_{\boldsymbol{\Theta}_{ij}^{(\cup)} | \mathcal{M}_{\cup}} P(\mathcal{D}_{\cup} | \boldsymbol{\Theta}_{ij}^{(\cup)}, \mathcal{M}_{\cup})
\end{aligned}
\tag{4.20}
$$

where $S_{i\eta}/T_{i\eta}$ is the Bayes factor for whether $\boldsymbol{\Theta}_{i\eta}^{(\cdot)}$, the parameter set that defines $X_i|\pi_{i\eta}$, is different across the two groups.

Table 2 shows the Bayes factors $S_{i\eta}/T_{i\eta}$ for the Balloons example. Bayes factors for *inflated* given $age = adult, act = dip$ and given $age = child, act = strech$ have values indicative of strong evidence for differences in distributions. Indeed, those are the combinations of values for which the two generating rules of **and** vs **or** yield different results: $age = adult, act = stretch$ is true for both 'adult **or** stretch' and 'adult **and** stretch,' and similarly, $age = child, act = dip$ is false for both rules. However, the configurations $age = adult, act = dip$ and given $age = child, act = stretch$ both yield true for the **or** rule and false for the **and** rule. This shows that the approach correctly identifies the parameter differences that explain the distribution-wide differences.

4.3.2 Detecting differences in partially similar models

One interesting and useful task is the detection of differences in distributions that have a mix of parameters that are different across groups and parameters that are shared among groups. The Bayes factor S/T in Equation (4.9), however, as stated above, compares the hypothesis of sharing no parameters across the groups, to the hypothesis of sharing all the parameters. Similarly, the Bayes factor S_i/T_i compares these models at the node level, meaning that it is a measure of whether *every* conditional distribution of X_i given its parents is different across the two groups. However, we are often interested in obtaining a measure that is sensitive to the presence of changes in only some conditional distributions of X_i, while other conditional distributions may indeed be identical across groups. This section derives a score

Table 2: Parameter-wise scores obtained for the Balloons data.

Variable X_i	$\lvert\pi_{i\eta}$	$S_{i\eta}/T_{i\eta}$
inflated	$\lvert age = adult, act = dip$	25.20
inflated	$\lvert age = adult, act = stretch$	0.2889
inflated	$\lvert age = child, act = dip$	0.2889
inflated	$\lvert age = child, act = stretch$	25.20
act	(no parents)	0.8076
age	(no parents)	0.8076
color	(no parents)	0.3754
size	(no parents)	0.3754

for identifying such differences, and while the details are presented in the context of the multi-model, they are equally applicable to the uni-model, since the latter is a special case of the former with additional structural constraints.

At the variable level, what we are really interested in is the posterior odds of seeing any difference between the groups in the conditional distribution of X_i given its parents. Let $p_{i\eta} = P(\mathbf{\Theta}_{i\eta}^{(1)} = \mathbf{\Theta}_{i\eta}^{(2)})$ denote the prior probability that the distribution of $X_i|\pi_{i\eta}$ is the same across the two groups. Using these priors and applying Bayes' rule we can derive a posterior probability that all parameters defining the distribution of X_i are the same for the two groups:

$$P(\mathbf{\Theta}_i^{(1)} = \mathbf{\Theta}_i^{(2)}|\mathcal{D}_1, \mathcal{D}_2) = \frac{\prod_{\eta=1}^{H_i} P(\mathcal{D}_1, \mathcal{D}_2|\mathbf{\Theta}_{i\eta}^{(1)} = \mathbf{\Theta}_{i\eta}^{(2)}) P(\mathbf{\Theta}_{i\eta}^{(1)} = \mathbf{\Theta}_{i\eta}^{(2)})}{P(\mathcal{D}_1, \mathcal{D}_2)} \tag{4.21}$$

where the numerator is $\prod_{\eta=1}^{H_i} T_{i\eta} p_{i\eta}$. Denote the power set of $\{1, \ldots, H_i\}$ by $\mathbb{P}(\{1, \ldots, H_i\})$.

The denominator of (4.21) can be expressed as follows:

$$\begin{aligned} P(\mathcal{D}_1, \mathcal{D}_2) \;=\; & \sum_{A \in \mathbb{P}(\{1,\ldots,H_i\})} \prod_{\eta \in A} \overbrace{P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} \perp \boldsymbol{\Theta}_{i\eta}^{(2)})}^{S_{i\eta}} \times \\ & \prod_{\eta \in \{1,\ldots,H_i\} \setminus A} \overbrace{P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)})}^{T_{i\eta}} \times \\ & \prod_{\eta \in A} \overbrace{\left(1 - P(\boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)})\right)}^{(1-p_{i\eta})} \times \\ & \prod_{\eta \in \{1,\ldots,H_i\} \setminus A} \overbrace{P(\boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)})}^{p_{i\eta}} = \\ = & \prod_{\eta=1}^{H_i} (S_{i\eta}(1 - p_{i\eta}) + T_{i\eta} p_{i\eta}) \;\; . \end{aligned} \tag{4.22}$$

Note at the η-level, only two cases are considered: either all $b\Theta_{i\eta}$ (all conditional distributions compatible with η) are different across the two groups, or all are the same. This is because η is defined to be the finest level at which the conditional distributions of the two models can be compared. From Equations (4.21) and (4.22) we can then obtain the posterior odds of a difference in any parameter of X_i:

$$\begin{aligned} O_i = \frac{1 - P(\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)} | \mathcal{D}_1, \mathcal{D}_2)}{P(\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)} | \mathcal{D}_1, \mathcal{D}_2)} = \frac{1}{P(\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)} | \mathcal{D}_1, \mathcal{D}_2)} - 1 = \\ = \frac{\prod_{\eta=1}^{H_i} (S_{i\eta}(1 - p_{i\eta}) + T_{i\eta} p_{i\eta})}{\prod_{\eta=1}^{H_i} T_{i\eta} p_{i\eta}} - 1 = \left(\prod_{\eta=1}^{H_i} \left(\frac{S_{i\eta}(1 - p_{i\eta})}{T_{i\eta} p_{i\eta}} + 1 \right) \right) - 1 \;\; . \end{aligned} \tag{4.23}$$

In the absence of information that would lead one to expect differences in some parameters more than in others, the priors $p_{i\eta}$ can be related to the prior probability p_i of seeing no difference in the conditional distribution of variable X_i by the relation $p_{i\eta} = p_i^{1/H_i}$.

The same approach can be applied to the entire data to obtain posterior odds of observing a difference anywhere in the network, expressed as

$$O = \frac{1 - P(\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)} | \mathcal{D}_1, \mathcal{D}_2)}{P(\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)} | \mathcal{D}_1, \mathcal{D}_2)} = \left(\prod_{i=1}^{n} \prod_{\eta=1}^{H_i} \left(\frac{S_{i\eta}(1 - p_{i\eta})}{T_{i\eta} p_{i\eta}} + 1 \right) \right) - 1 \;\; . \tag{4.24}$$

Table 3: Variable-wise Bayes factors and posterior odds obtained for the Balloons data.

Variable	S_i/T_i	O_i
inflated	53.00	36.01
act	0.8076	0.8076
age	0.8076	0.8076
color	0.3754	0.3754
size	0.3754	0.3754

Using (4.24) entails that the prior for seeing no difference between the two groups is $p = \prod_{i=1}^{n} \prod_{\eta=1}^{H_i} p_{i\eta}$. Given such an overall prior p, a natural choice for noninformative priors is $p_{i\eta} = p^{1/(nH_i)}$: this choice of priors assumes that we are equally and independently likely to see a difference in each variable, and equally and independently likely to see a change in each conditional probability distribution of each variable.

4.3.3 Examples

I applied this approach to the Balloons example with priors set in this manner with all $p_i = 1/2$. Computing the model-wise posterior odds for the Balloons example yields $O = 227.8$. As a point of comparison, the model-wise Bayes factor is 4.872. This is not surprising since in the Balloons example, the generating model is partially different between the groups, and we expect the posterior odds to reliably detect this event. Table 3 shows the posterior odds O_i obtained for the Balloons example with alongside the Bayes factors S_i/T_i from Table 1. For this particular prior choice, $O_i = S_i/T_i$ for variables that have no parents. In general, when $H_i = 1$, that is, when the intersection of the parent sets of X_i from the three networks is empty, $O_i = S_i(1 - p_i)/(T_i p_i)$. We see that the O_i score correctly identifies the variable *inflated* as the variable changed.

To illustrate what happens to the relationship between O_i and S_i/T_i as H_i increases,

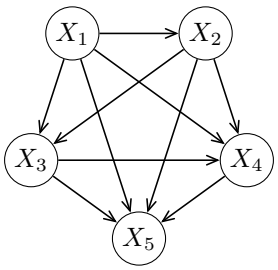

Figure 6: A network structure example.

consider the network in Figure 6, with probability distributions defined as follows:[1]

$$P(x_{ik_i}|\{x_{\ell k_\ell}|\ell \in \mathbb{N}, \ell < i\}) = \begin{cases} 0.7 & \text{if } \bigoplus_{\ell=1}^{i} x_{\ell k_\ell} = T \\ 0.3 & \text{otherwise.} \end{cases} \tag{4.25}$$

Table 4 gives a summary of a 2000-point dataset generated from this network in terms of the number of times each possible configuration of the variables appears. For the purposes of this example, we compare this group to a copy of itself, meaning that $\mathcal{D}_1$ and $\mathcal{D}_2$ each contain the data in Table 4. The models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_\cup$ are all correct recoveries of the network structure in Figure 6. Table 5 compares the scores S_i/T_i obtained for each variable to the scores O_i obtained using Equation (4.23) with $p_i = 1/2$.

We can see that as the number of parents of X_i increases, S_i/T_i exponentially decreases for data that shows no difference between the two groups. This is to be expected, since $S_{i\eta}/T_{i\eta}$ is always less than one for data that shows no change, and since H_i is exponential in the number of parents of X_i, more such terms are multiplied. This effect is detrimental to using the Bayes factor S_i/T_i as a measure for detecting a difference in a variable. To illustrate, consider the same network from Figure 6 with the same parameters from (4.25) with the exception that one conditional probability distribution is changed to:

$$P(X_5|X_1 = T, X_2 = T, X_3 = F, X_4 = F) = \begin{cases} 0.9 & \text{for } X_5 = F \\ 0.1 & \text{for } X_5 = T \end{cases} \tag{4.26}$$

[1] $\mathbb{N}$ is the natural numbers $1, 2, 3, \ldots$ and $\bigoplus$ is the 'exclusive or' operator. $X_1, \ldots, X_5$ are binary variables taking true/false values.

Table 4: Summary of 2000-point dataset generated from the network in Figure 6.

			X_4	F		T	
X_1	X_2	X_3	X_5	F	T	F	T
F	F	F		3	15	22	9
F	F	T		74	22	8	27
F	T	F		156	63	26	70
F	T	T		12	15	74	34
T	F	F		351	118	58	146
T	F	T		23	67	143	55
T	T	F		8	26	62	27
T	T	T		140	57	23	66

Table 5: Variable-wise scores obtained for comparing the data in Table 4 ($\mathcal{D}_1$) to a copy of that data ($\mathcal{D}_2$).

Variable	S_i/T_i	O_i
X_1	3.680×10^{-2}	3.680×10^{-2}
X_2	2.790×10^{-3}	4.500×10^{-2}
X_3	3.750×10^{-5}	6.306×10^{-2}
X_4	2.995×10^{-8}	9.267×10^{-2}
X_5	3.093×10^{-13}	1.353×10^{-1}

Table 6: Summary of the 2000-point dataset generated with the perturbed conditional distribution from (4.26).

			X_4	F		T	
X_1	X_2	X_3	X_5	F	T	F	T
F	F	F		4	7	32	8
F	F	T		63	26	12	30
F	T	F		152	60	32	55
F	T	T		10	26	56	26
T	F	F		322	162	69	151
T	F	T		27	69	127	63
T	T	F		27	4	68	22
T	T	T		134	56	27	73

Table 6 shows a summary of a group of 2000 data points generated from this perturbed version of the network. Let $\mathcal{D}_1$ be the previous data, from Table 4, and let $\mathcal{D}_2$ be the data from Table 6, generated from the perturbed network. When comparing these two groups, the models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_\cup$ again recover the network structure of Figure 6 correctly, and the scores shown in Table 7 are obtained.

We can see from Table 7 what happens with the scores and why S_i/T_i is not a good indicator for determining which variable was changed, since even though X_5 is the only variable the distribution of which is different across the two groups, S_5/T_5 is close to (and less than) S_2/T_2 and is less than S_1/T_1 by orders of magnitude. This is because, even though the $S_{5\eta}/T_{5\eta}$ term in the product that corresponds to the difference introduced in (4.26) is large (it is 310,400), its contribution to the product is drowned out by the other 15 $S_{5\eta}/T_{5\eta}$ terms (14 of which are between 0.1 and 0.7). The posterior odds O_i avoids this problem; the posterior odds for X_5 are much greater than 1, while the posterior odds for each of the rest of the variables are below 1.

Table 7: Variable-wise scores obtained for comparing the data in Table 4 to the data in Table 6.

Variable	S_i/T_i	O_i
X_1	6.425×10^{-2}	6.425×10^{-2}
X_2	3.769×10^{-3}	5.383×10^{-2}
X_3	5.152×10^{-5}	6.773×10^{-2}
X_4	2.667×10^{-6}	2.487×10^{-1}
X_5	3.766×10^{-3}	1.936×10^{5}

In general, variable-level posterior odds score O_i is expected to be better than the Bayes factor S_i/T_i for detection of differences where we have a mix between parameters defining the distribution of X_i that are and are not different between the two groups. If we are instead interested in finding out whether all parameters associated with a variable X_i are different, the variable-level Bayes factor is expected to be the better metric. At the parameter ($\mathbf{\Theta}_{i\eta}$) level, the posterior odds score is more general than the Bayes factor, since the posterior odds is just the Bayes factor multiplied by the prior odds $(1-p_{i\eta})/p_{i\eta}$ of a difference in $\mathbf{\Theta}_{i\eta}$, hence with a choice of $p_{i\eta} = 1/2$, the two scores are equivalent. In general, the choice of priors will affect the posterior odds, and the score will be best at detection when priors are close to the distribution of the differences in the data, and may perform poorly when priors are vastly different from the distribution of differences in the data.

4.3.4 Model synthesis for clinical difference detection

The posterior odds serves to find statistically significant differences. Having detected a statistically significant difference between two sets of parameters $\mathbf{\Theta}_{i\eta}^{(1)}$ and $\mathbf{\Theta}_{i\eta}^{(2)}$, in order to find clinically significant differences we must match up the appropriate parameters and compare them to evaluate the size of the difference. It turns out that there is a straightforward method for synthesizing a single BN model from the $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ using the results of

the statistical detection test. Once we have a single model, testing for clinical significance becomes analogous to the uni-model case.

The synthesized model will contain the group indicator variable Z. The distribution of Z can be easily estimated directly from the data. For variables that show no statistically significant differences across the entire node (the posterior odds of observing a difference in parameters is one), the variable keeps its parents and parameter distribution from $\mathcal{M}_\cup$. For variables that are found to have a statistically different distribution, we group the parameters by ηs: if the posterior of $\mathbf{\Theta}_{i\eta}$ is found not to be statistically different, that group of parameters is inherited from the $\mathcal{M}_\cup$ network (and hence the X_i inherits the parents from $\mathcal{M}_\cup$). For parameter groups η where a statistical difference is found, X_i receives an incoming arc from Z, and inherits the parents (and parameter distributions) from $\mathcal{M}_1$ and $\mathcal{M}_2$ for the conditional distributions associated with $Z = 1$ and 2 respectively. To summarize, the parents of X_i in the synthesized network are

$$\Pi_i = \begin{cases} \text{if } O_i \leq 1 \text{ then} & \Pi_i^{(\cup)} \\ \text{otherwise} & \bigcup_{\eta=1}^{H_i} \begin{cases} \Pi_i^{(\cup)} & \text{if } \dfrac{S_{i\eta}(1-p_{i\eta})}{T_{i\eta}p_{i\eta}} \leq 1 \\ \{Z\} \cup \Pi_i^{(1)} \cup \Pi_i^{(2)} & \text{otherwise.} \end{cases} \end{cases} \tag{4.27}$$

If $O_i \leq 1$, the parameter distribution of X_i is simply the parameter distribution of X_i in $\mathbf{M}_\cup$. Recall that η corresponds to a variable assignment of $\overset{\cap}{\Pi}_i = \Pi_i^{(1)} \cap \Pi_i^{(2)} \cap \Pi_i^{(\cup)}$. Hence $\overset{\cap}{\Pi}_i$ is a subset of Π_i from (4.27), and each variable assignment π_{ij} matches exactly one η (but each η may match many different js). Denoting the η that matches a j by $\eta(j)$ we can express the distribution of X_i when $O_i > 1$ as:

$$P(X_i|\pi_{ij}, z) \equiv \begin{cases} P(X_i|\pi_{ij}^{(\cup)}, \mathcal{M}_\cup) & \text{if } \dfrac{S_{i\eta(j)}(1-p_{i\eta(j)})}{T_{i\eta(j)}p_{i\eta(j)}} \leq 1 \\ P(X_i|\pi_{ij}^{(z)}, \mathcal{M}_z) & \text{otherwise.} \end{cases} \tag{4.28}$$

Consider the example in Figure 7, where in $\mathcal{M}_1$ (Figure 7a) X_1 and X_2 are parents of X_4, in $\mathcal{M}_2$ (Figure 7b) only X_1 is a parent of X_4, and in $\mathcal{M}_\cup$ (Figure 7c) X_1 and X_3 are parents of X_4. Suppose that the variables are binary. Table 8 shows how the various combinations of value assignments for X_1, X_2, and X_3 map to the η-indexes and j-indexes in each model

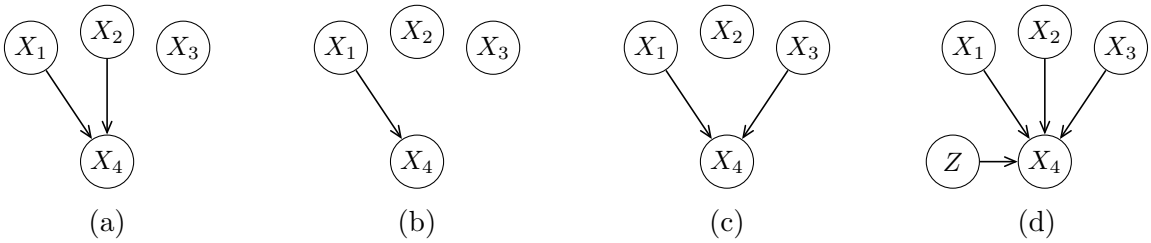

Figure 7: An example of three models (a) $\mathcal{M}_1$, (b) $\mathcal{M}_2$, and (c) $\mathcal{M}_\cup$, as well as (d) a BN synthesized from the three models.

Table 8: Mapping of indexes for the network synthesis example.

X_1	X_2	X_3	$j^{(1)}$	$j^{(2)}$	$j^{(\cup)}$	η
x_{11}	x_{21}	x_{31}	1	1	1	1
		x_{32}	1	1	2	
	x_{22}	x_{31}	2	1	1	
		x_{32}	2	1	2	
x_{12}	x_{21}	x_{31}	3	2	3	2
		x_{32}	3	2	4	
	x_{22}	x_{31}	4	2	3	
		x_{32}	4	2	4	

Table 9: The conditional probability distribution of X_4 in the synthesized network in the network synthesis example. An asterisk in a variable's column indicates that the probability does not depend on the value of the variable.

η	X_1	X_2	X_3	Z	$P(X_4 \mid X_1, X_2, X_3, Z)$
1	x_{11}	x_{21}	*	1	$P(X_4 \mid x_{11}, x_{21}, \mathcal{M}_1)$
		x_{22}			$P(X_4 \mid x_{11}, x_{22}, \mathcal{M}_1)$
		*	*	2	$P(X_4 \mid x_{11}, \mathcal{M}_2)$
2	x_{12}	*	x_{31}	*	$P(X_4 \mid x_{12}, x_{31}, \mathcal{M}_\cup)$
			x_{32}		$P(X_4 \mid x_{12}, x_{32}, \mathcal{M}_\cup)$

for X_4. Furthermore, suppose that the posterior odds for X_4 exceed one ($O_4 > 1$), and that $\frac{S_{41}(1-p_{41})}{T_{41}p_{41}} > 1$ and $\frac{S_{42}(1-p_{42})}{T_{42}p_{42}} \leq 1$. Based on these values for the variable and parameter-group posteriors, (4.27) gives the network structure in Figure 7d, and (4.28) and conditional probability distribution in Table 9.

The result of this network synthesis process is a BN with context-specific independence. Context-specific independencies in BNs are independencies that are additional to the independence captured by the network structure. Boutilier et al. (1996) define local context-specific independence in Bayesian networks as a conditional independence between a variable X_i and a subset $\mathbf{A} \subset \Pi_i$ of its parents given an assignment $\mathbf{b}$ of a subset $B \in \Pi_i$ of the node's remaining parents. This sort of independence cannot be represented in the BN structure when the conditional independence does not hold for some different value assignment $\mathbf{b}'$ of B. This is the sort of independence that is created by the synthesis process when $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ have different parents. Some instances of local contextual independence that hold in the example above are $(X_4 \perp X_3 | x_{11}, x_{21}, z = 1), (X_4 \perp X_, (X_4 \perp Z | x_{12}, x_{31})$, and $(X_4 \perp X_2 | x_{12}, X_{32})$.

By the process above we construct a single BN that captures the statistical differences between the groups that are identified as significant according to the posterior odds. Having

a single BN that captures the differences allows us to apply tests of clinical significance in the exact same manner we would for the uni-model approach. We can also apply the explanation methods developed in Chapter 6 to this multi-model constructed BN. The case study in Section 6.4.1 is an example.

4.4 LIST OF CLINICAL SIGNIFICANCE TESTS

Having a method to obtain a single probabilistic model using either uni-model or the multi-model approach, we can in principle test the clinical significance not only of parameter differences, but of any event e described in terms of an assignment of BN variables to values, optionally conditioned on an assignment of other BN variables. When e is $x_{ik}|\pi_{ij}$, we test the difference in the parameter θ_{ijk}, but we can also test an event x_{ik}, π_{ij} (is a joint assignment of a variable and its parents clinically significantly different?) or an event x_{ik} (is the marginal probability of a variable taking a value clinically significantly different?)

For an event e, I look at two measures of effect size that quantify how different the probability of e is between the groups: the difference and the ratio between the probabilities in each group. Let $P_{\mathbf{X}|\boldsymbol{\theta}}(e|z)$ denote the probability of event e in group z as computed by a BN with parameters $\boldsymbol{\theta}$. As we have distributions over the parameters $\boldsymbol{\Theta}$ of the BN, the posterior distributions obtained from the data and Dirichlet priors, the probability of an event e in group z given the random vector of parameters $\boldsymbol{\Theta}$ as a random variable $P_{\mathbf{X}|\boldsymbol{\Theta}}(e|z)$. To get a point estimate of this probability, we can marginalize away $\boldsymbol{\Theta}$ to obtain $P_{\mathbf{X}}(e|z) = E_{\boldsymbol{\Theta}}P_{\mathbf{X}|\boldsymbol{\Theta}}(e|z)$. Using this notation we can express the following clinical significance tests:

1. Absolute difference test:

$$|P_{\mathbf{X}}(e|Z=2) - P_{\mathbf{X}}(e|Z=1)| > \delta \tag{4.29}$$

2. Probabilistic absolute difference test:

$$P_{\boldsymbol{\Theta}}(|P_{\mathbf{X}|\boldsymbol{\Theta}}(e|Z=2) - P_{\mathbf{X}|\boldsymbol{\Theta}}(e|Z=1)| > \delta) > 1-\alpha \tag{4.30}$$

3. Absolute log-ratio test:

$$|\log(P_{\mathbf{X}}(e|Z=2)) - \log(P_{\mathbf{X}}(e|Z=1))| > \epsilon \tag{4.31}$$

4. Probabilistic absolute log-ratio test:

$$P_{\mathbf{\Theta}}(|\log(P_{\mathbf{X}|\mathbf{\Theta}}(e|Z=2)) - \log(P_{\mathbf{X}|\mathbf{\Theta}}(e|Z=1))| > \epsilon) > 1-\alpha \tag{4.32}$$

Given a Bayesian network, the absolute difference and absolute log-ratio tests for any event e are simply computed by BN inference on the network. For example, when the event e is x_{ik}, the absolute difference test is

$$|P_{\mathbf{X}}(x_{ik}|Z=2) - P_{\mathbf{X}}(x_{ik}|Z=1)| > \delta \tag{4.33}$$

or, when the event e is $x_{ik}|\pi_{ij}$, the log-ratio test is

$$|\log(P_{\mathbf{X}}(x_{ik}|\pi_{ij}, Z=2)) - \log(P_{\mathbf{X}}(x_{ik}|\pi_{ij}, Z=1))| > \epsilon \ , \tag{4.34}$$

which one may also write as $|\log\theta_{ijk}^{(2)} - \log\theta_{ijk}^{(1)}| > \epsilon$, the test in (4.6).

The probabilistic absolute difference test and the probabilistic log-quotient test are evaluated with respect to a fixed network structure and posterior distributions of network parameters. I approximate the posterior Dirichlet distributions of the parameters by sampling. Parameter independence allows us to treat and sample the network parameters as independent Dirichlet random variables. In this manner I can generate a collection of BNs with parameters sampled from the corresponding Dirichlet distributions, all with the same structure.

The probabilistic tests have a common structure: there is an inner test, that corresponds to one of the absolute tests, and an outer test of the probability of passing the inner test. Given a sample of many BNs, I can then use BN inference in each network to perform the inner test, and estimate the probability needed for the outer test by counting the proportion of BNs for which the inner test passed.

Note that when the event e is $x_{ik}|\pi_{ij}$, we can avoid performing BN inference, and can hence avoid sampling parameters for the entire network. Since the test in that case is comparing only a pair of parameters, we need only sample those compared parameters from their corresponding Dirichlet distributions.

5.0 AN EMPIRICAL EVALUATION OF THE DETECTION OF DIFFERENCES IN DISTRIBUTIONS

The aim of this empirical evaluation is to test the difference recovery hypothesis of Section 4.1.1. I ran an array of experiments that compare data generated from pairs of BNs that have known parametric differences (a conditional distribution $X_i|\pi_{ij}$ is different) or structural differences (an arc that is present in one BN is absent in the other). I focus on quality of detection of variable-level differences, since that is the level of granularity at which successful detection of a difference is well-defined for both the parametric and the structural perturbations. The following section describes the data and the experimental setup in detail. It is followed by the evaluations of the statistical difference detection methods and of the clinical difference detection tests.

5.1 DATA AND EXPERIMENTAL SETUP

Since in real-world data the differences between groups of data are not known in advance, for the evaluation I generated pairs of data groups from known distributions that are based on real-world data. I chose to learn networks from which to generate data because publicly available BN models are overwhelmingly diagnostic, meaning that only a handful of variables in the network are intended to be observed and the relationships between them are mediated by variables that are intended to be hidden, whereas I would like to have a ground-truth model that directly relates observed variables to each other. I picked data where all variables are categorical, since the BD score is designed for BNs that represent multinomial distributions. In this evaluation I used several datasets available from the UCI

Table 10: Description of data used.

Set	# variables	# values		
		Min.	Med.	Max.
balance-scale	5	3	5	5
car	7	3	3	4
hayes-roth	5	3	4	4
nursery	9	2	3	5

Machine Learning Repository (Bache and Lichman, 2013). Table 10 lists the datasets and provides brief descriptions in terms the number of variables in the data, and the minimum, median, and maximum number of values per variable. I learned a BN from the data for each of these sets, which is referred to as the "original BN" in the following description of the data-generation process.

I ran blocks of tests, where each block is characterized by a data source (one of the UCI datasets), a type of perturbation, the number of perturbations, and the number of samples per group. Each block consists of 20 group pairs, where each pair consists of a group of points generated from the original BN of the data source and a group of points generated from a perturbed BN of a data source (a different perturbed BN is obtained for each group pair). The original BN for a data source is the one mentioned above, learned directly from the original data. The perturbed BN was obtained by performing perturbations to the original BN. There are two possible categories of perturbations: parametric perturbations and structural perturbations. A parametric perturbation was performed by uniformly randomly selecting a variable X_i to perturb and then selecting a conditional distribution $X_i|\pi_{ij}$ to perturb, and then replacing its probability mass vector with a random non-trivial permutation of itself. For example, if the probability mass vector of $X_i|\pi_{ij}$ was $(0.2, 0.5, 0.3)$, a possible perturbed probability mass vector might be $(0.5, 0.2, 0.3)$. The trivial permutation would be $(0.2, 0.5, 0.3)$, and we would not select it as a possible permutation. A structural perturbation

was performed by randomly (with probability 1/2) deciding whether to remove or add an arc, and then selecting a random arc to add (or remove) from the existing (or absent) arcs in the network. A node (variable) is considered perturbed by a structural perturbation only if an arc into the node is added or removed.

I provide the ordering of the variables in the generating model to the logistic regression method (see Section 5.2.1) so that it may take advantage of that information to improve computational efficiency and detection performance. For the difference detection methods of Chapter 4, I show results obtained both with and without using the ordering information provided to the introduced methods. It is expected that ordering information would improve detection quality, however, it is also of interest to evaluate the detection quality without knowledge of the variable ordering, since we often do not know the ordering perfectly when we examine real-world data.

5.2 EVALUATION OF STATISTICAL SIGNIFICANCE TESTS

I evaluated the performances of the uni-model variable selection and of the multi-model posterior odds O_i in Equation 4.23 as a score for detecting variable-level differences. I compared the performance of both scores with and without variable ordering information to the baseline method which I describe below.

5.2.1 Baseline Method

As a point of comparison for the statistical difference detection methods, I chose to simulate a process often followed by analysts, statisticians, and researchers, where logistic regression models with interactions are constructed to predict a variable X_i using candidate predictors, and the researcher would judge a predictor's relevance based on the strength of its corresponding weight.

I detect variable-wise differences across two pre-defined groups using lasso-regularized logistic regression. To do so, I add the group indicator Z to the data. Lasso regularization

solves

$$\min_{\beta_0,\beta} \frac{1}{N} Deviance(\beta_0, \beta) + \lambda \sum_{i=1}^{p} |\beta_i| \tag{5.1}$$

for a model with p predictors where β_0 and β are the weights corresponding respectively to the constant term and the predictors in the logistic model, and λ is the regularization parameter. To detect differences across groups, I use this model for predicting each variable X_i given all the other data variables $X_j : j \in \{1, \ldots, i-1\}$ that precede it in the variable ordering, the group indicator Z, and interactions of Z with each of the data variables X_j. I handle more-than-binary X_i by using multinomial logistic regression, and I handle more-than-binary X_j's by binary-coding them.

The effect of regularization is that as λ decreases from $+\infty$, predictors enter the model (β terms change from zero to nonzero). The largest value of λ at which a given predictor becomes nonzero can then be used as a score of how useful that predictor is for predicting X_i. Hence, for each X_i we can use the largest λ that corresponds to a nonzero β value for Z or an interaction with Z as the score for seeing a difference in the distribution of X_i across groups.

5.2.2 Results

First, let us compare the fitness of the Bayes factor as compared to the posterior odds at the detection of statistical differences. Tables 11, 12, 13, and 14 show the areas under the ROC curves (AUCs) for statistical difference recovery on the 72 blocks of tests using the Bayes factor and using the posterior odds for the uni-model approach.

Next, tables 15, 16, 17, 18, 19, 20, 21, and 22 show the AUCs for statistical difference recovery using logistic regression approach, and using the posterior odds with the uni model and the multi-model.

AUCs were computed by obtaining ROC curves based on a score. For each pair of data groups compared, each method gave a score for each variable in the data. For the posterior odds scores, the posterior odds O_i with a prior of $p_i = 1/2$ was used. For the uni-model Bayes factor scores, the ratio of the marginal likelihood of including Z as a parent of the variable X_i or not including it was used. For the logistic regression approach, the λ described

above was used as the score. The semantics of obtaining an ROC curve from a score are that all cases that fall below a certain score are identified as negative, and those that fall above it are identified as positive. The ROC curve is then a curve of plotting the proportion of true positive cases against the proportion of false negative cases for the full range of scores that appear in the data (Fawcett, 2004). Each variable in each pair of groups (there are 20 pairs in each block) constitutes a "case" for the purposes of the ROC. The gold standard for each case is whether the variable was perturbed when generating that particular pair of groups.

Each table corresponds to a UCI Repository data source and is structured as follows: The first column indicates the perturbation type (structural or parametric), the second column indicates the number of perturbations, and the third column indicates the number of data points per group used in each test. For comparing the Bayes factor to the posterior odds, Tables 11, 12, 13, and 14 have two groups of columns, results when no variable ordering information is provided to the tests on the left and results when variable information is provided on the right. For comparing the logistic regression method, the uni-model, and the multi-model, Tables 15, 17, 19, and 21 show the results when no variable ordering information is provided to the tests from Chapter 4, while tables 16, 18, 20, and 22 show the results when the true variable ordering is provided. The logistic regression method is provided with the true variable ordering in all cases. The AUCs for the the logistic regression method (indicated by λ), the uni-model detection method (Uni/U), and the multi-model detection method (Multi/M) were compared. The tables also show the p-values for two-tailed tests of the difference between the AUCs of each pair of methods, based on (DeLong et al., 1988). The p-values that are significant at $\alpha = 0.05$ are displayed in bold.

The results show that generally, when there are less data points per group and more variables in the data, the detection performance is lower. This is in part because the models use structure learning, since it is more difficult to recover a structure accurately when there is less data and when the structure to recover is more complex (due to more variables). The typical issues of attempting to make inferences about a distribution from small samples apply as well. Also note that all methods are generally better at detecting structural differences than parametric ones. This is because a structural difference reflects a more substantial distributional difference than a simple parametric one, since it can be expressed as a collection

of parametric differences in a network containing the removed or added arcs.

The comparison of the Bayes factor to the posterior odds confirms that, as expected, the posterior odds is better suited for this task, since we are interested in detecting differences in similar models. The Bayes factor performs well, and sometimes better, than the posterior odds at detecting structural perturbations. This is likely because a structural perturbation constitutes a change in all parameters associated with a variable.

In comparing the uni-model, multi-model, and logistic regression methods, overall, the experiments show consistently good AUCs for the multi-model score over the various generated group pairs. Of the 72 blocks of tests, when the models are not given ordering information, the multi-model method performed better than the uni-model method in 28 blocks, with none of the differences being significant at the $\alpha = 0.05$ level, and the uni-model performed better than the multi-model in 34 blocks, with only one difference being significant at the $\alpha = 0.05$ level. The two models had equal performance in 10 blocks. When the models are given ordering information, the multi-model method performed better than the uni-model method in 26 blocks, with four of the differences being significant at the $\alpha = 0.05$ level, and the uni-model performed better in 18 blocks, with only two differences being significant at the $\alpha = 0.05$ level. Given the ordering, the two models had equal performance in 28 blocks.

As expected, higher AUCs are obtained when the multi-model and the uni-model use ordering information. Interestingly, Tables 15 and 16 are almost identical. It appears that this happens because the correct variable ordering is recovered in the model construction phase for the data generated based on the *balance-scale* dataset.

Compared to the logistic regression baseline, the multi-model AUC is higher in 52 of the 72 test blocks when no ordering information is provided, and it is higher in 61 of the 72 test blocks when ordering information is provided. At the $\alpha = 0.05$ significance level, without ordering information, the multi-model performed statistically significantly better than the logistic regression baseline in 21 test blocks, and statistically significantly worse in only 8 test blocks. With ordering information, the multi-model performed statistically significantly better in 34 test blocks, and worse in only one block.

The uni-model AUC is higher than the logistic regression AUC in 48 of the 72 test blocks when no ordering information is provided, and in 55 test blocks when ordering information

is provided. At the $\alpha = 0.05$ significance level, without ordering information, the uni-model performed statistically significantly better than the logistic regression baseline in 23 test blocks, and statistically significantly worse in 9 test blocks. With ordering information, the uni-model performed statistically significantly better in 33 test blocks, and worse in 3 blocks.

Overall, the multi-model approaches gives the best results, with the uni-model approach performing almost as well, with both methods performing better than logistic regression on both tests. The very close performance of the multi- and uni-models is likely a result of learning very similar networks. For any variable for which the multi-model learns the same parents in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$, the two scores are equivalent, which explains why we see these models get identical AUCs for many of the parametric perturbations. The uni-model tends to perform better or as well as the multi-model in most parametric tests. This is likely because the multi-model incorrectly learns a structural difference where there is none in those cases where it performs worse on parametric perturbation detection.

The cases in which the uni- and multi-model approaches underperform relative to logistic regression are those where they are not provided an ordering and the perturbations are structural. One possible explanation is that when a structure is learned with the wrong ordering, the multi-model approach identifies that a difference in the dependence between two variables exists, but attributes it to the wrong variable. For example, if the generating model for one group is $X_a \quad X_b$ (no arc) and for another is $X_a \rightarrow X_b$, but the model learned for the latter group is $X_a \leftarrow X_b$ (which can still accurately model the dependence relationship in the statistical sense), the detected variable-level difference would be attributed to X_a, but the gold standard difference would be considered a difference in X_b, counting against the detection. The fact that the logistic regression method is provided with the correct ordering prevents it from making this mistake.

In Tables 17, 19, 21 and 22, when logistic regression performs better, it often does so when there are smaller samples. This may possibly be explained by the fact that logistic regression does not build a full conditional probability table for each node, and the number of parameters it fits is linear in the number of predictors, while the other methods which use BNs, fit a full multinomial conditional probability table, the size of which is exponential in the number of predictors. Since a model with less parameters can be more reliably fit

given less samples, this difference in the number of parameters may account for the better performance of logistic regression on smaller sample sizes.

In spite of underperformance in detecting structural perturbations without ordering information, the uni- and multi-model approaches performs better than the logistic regression method on average when they are provided with ordering information. The uni- and multi-model approach are much better than logistic regression at the detection of parametric perturbations.

Table 11: Statistical difference detection AUCs comparing the Bayes factor to the posterior odds score under the uni-model approach on the *balance-scale* data.

			Without ordering			With ordering		
			BF AUC	PO AUC	p-value	BF AUC	PO AUC	p-value
Parametric perturbations	1	500	0.8431	0.8931	0.1382	0.8431	0.8931	0.1382
		1000	0.8925	0.9444	0.1229	0.8925	0.9444	0.1229
		5000	0.9175	0.9563	0.1846	0.9175	0.9563	0.1846
	3	500	0.7431	0.8099	0.1760	0.7483	0.8295	0.0781
		1000	0.7743	0.7963	0.5473	0.7747	0.8043	0.3999
		5000	0.8291	0.8936	**0.0112**	0.8291	0.8936	**0.0112**
	5	500	0.8224	0.8988	0.0596	0.8264	0.9142	**0.0205**
		1000	0.8433	0.8780	0.3558	0.8433	0.8780	0.3558
		5000	0.8571	0.8973	0.1882	0.8571	0.8973	0.1882
Structural perturbations	1	500	0.9712	0.9650	0.8474	0.9712	0.9669	0.8941
		1000	0.9575	0.9706	0.3385	0.9575	0.9706	0.3385
		5000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	3	500	0.9555	0.9800	0.0897	0.9555	0.9800	0.0897
		1000	0.9555	0.9840	0.1377	0.9555	0.9840	0.1377
		5000	0.9916	0.9992	0.3250	0.9916	0.9992	0.3250
	5	500	0.9333	0.9571	0.2926	0.9333	0.9571	0.2926
		1000	0.9487	0.9908	**0.0263**	0.9487	0.9908	**0.0263**
		5000	0.9829	0.9921	0.1907	0.9829	0.9921	0.1907

Table 12: Statistical difference detection AUCs comparing the Bayes factor to the posterior odds score under the uni-model approach on the *car* data.

			Without ordering			With ordering		
			BF AUC	PO AUC	p-value	BF AUC	PO AUC	p-value
Parametric perturbations	1	500	0.6554	0.7058	0.6540	0.6458	0.6954	0.6627
		1000	0.7258	0.7183	0.9415	0.7017	0.7212	0.8562
		5000	0.7579	0.7308	0.7944	0.7771	0.7188	0.5284
	3	500	0.7126	0.7887	0.2468	0.7297	0.7797	0.4326
		1000	0.7574	0.8180	0.3211	0.7587	0.8193	0.3149
		5000	0.8309	0.8291	0.9731	0.7955	0.8368	0.4590
	5	500	0.6656	0.7471	0.1583	0.7020	0.7626	0.3013
		1000	0.6532	0.7638	0.0686	0.7089	0.8038	0.1222
		5000	0.7430	0.7887	0.3581	0.7728	0.8447	0.1984
Structural perturbations	1	500	0.8954	0.9250	0.5971	0.9008	0.9308	0.5956
		1000	0.8833	0.9463	0.1666	0.9158	0.9521	0.3455
		5000	0.9475	0.9804	0.3250	0.9533	0.9854	0.3185
	3	500	0.9128	0.8263	0.0505	0.9253	0.8621	0.1435
		1000	0.9214	0.8872	0.2866	0.9506	0.9368	0.6459
		5000	0.9549	0.9497	0.7820	0.9832	1.0000	0.1597
	5	500	0.9095	0.8164	**0.0046**	0.9371	0.8768	0.0709
		1000	0.9191	0.8881	0.1402	0.9483	0.9487	0.9871
		5000	0.9063	0.9079	0.7851	0.9763	0.9961	0.1577

Table 13: Statistical difference detection AUCs comparing the Bayes factor to the posterior odds score under the uni-model approach on the *hayes-roth* data.

			Without ordering			With ordering		
			BF AUC	PO AUC	p-value	BF AUC	PO AUC	p-value
Parametric perturbations	1	500	0.5525	0.7244	0.0997	0.6456	0.7844	0.2083
		1000	0.5700	0.7625	0.0720	0.7231	0.8194	0.3405
		5000	0.5888	0.6713	0.3952	0.6663	0.8006	0.2318
	3	500	0.5925	0.7768	**0.0074**	0.7306	0.8462	0.0827
		1000	0.5990	0.7202	0.0671	0.7098	0.8422	0.0506
		5000	0.6246	0.6925	0.1726	0.7407	0.8515	0.0952
	5	500	0.5833	0.7183	**0.0484**	0.7098	0.8482	**0.0441**
		1000	0.6518	0.7371	0.2629	0.7217	0.8591	**0.0469**
		5000	0.6766	0.7326	0.2205	0.7143	0.8621	**0.0220**
Structural perturbations	1	500	0.6531	0.8219	**0.0472**	0.9106	0.9137	0.9564
		1000	0.7500	0.8156	0.3298	0.9269	0.9044	0.6487
		5000	0.8619	0.8425	0.5360	0.9644	0.9475	0.6434
	3	500	0.8498	0.9103	0.1211	0.9267	0.9111	0.6280
		1000	0.8918	0.9038	0.6816	0.9487	0.9131	0.2114
		5000	0.8894	0.8978	0.4924	0.9808	0.9411	**0.0203**
	5	500	0.8555	0.9049	0.0544	0.9440	0.9405	0.8449
		1000	0.9119	0.9288	0.4366	0.9457	0.9418	0.8098
		5000	0.9119	0.9128	0.6391	0.9549	0.9497	0.7490

Table 14: Statistical difference detection AUCs comparing the Bayes factor to the posterior odds score under the uni-model approach on the *nursery* data.

			Without ordering			With ordering		
			BF AUC	PO AUC	p-value	BF AUC	PO AUC	p-value
Parametric perturbations	1	500	0.5425	0.7369	0.0824	0.5269	0.7291	0.0720
		1000	0.5659	0.7388	0.0565	0.5697	0.7453	0.0543
		5000	0.5372	0.8641	**0.0009**	0.5372	0.8641	**0.0009**
	3	500	0.5846	0.6503	0.3238	0.6108	0.6552	0.4944
		1000	0.6179	0.6944	0.2484	0.6391	0.7023	0.3441
		5000	0.6782	0.7774	0.0789	0.6852	0.7896	0.0646
	5	500	0.5561	0.6479	0.1396	0.5646	0.6413	0.2092
		1000	0.5483	0.6655	0.0617	0.5773	0.6815	0.0907
		5000	0.5769	0.7293	**0.0068**	0.5919	0.7310	**0.0127**
Structural perturbations	1	500	0.7331	0.6528	0.4082	0.7406	0.6538	0.3529
		1000	0.6678	0.7688	0.2595	0.6850	0.7647	0.3166
		5000	0.7919	0.8678	0.3200	0.7928	0.8700	0.3115
	3	500	0.6788	0.7329	0.3684	0.6794	0.7220	0.4806
		1000	0.7734	0.7900	0.7526	0.7754	0.7937	0.7309
		5000	0.8966	0.8967	0.9970	0.9089	0.9028	0.8748
	5	500	0.8157	0.7745	0.3046	0.8279	0.7740	0.1772
		1000	0.8323	0.7622	0.1125	0.8323	0.7622	0.1125
		5000	0.9193	0.9351	0.4895	0.9345	0.9418	0.7314

Table 15: Statistical difference detection AUCs comparing the uni-model (using no variable ordering information), multi-model (using no variable ordering information), and logistic regression (using variable ordering information) approaches on tests using the *balance-scale* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.7031	0.8931	0.8931	**0.0074**	**0.0074**	1.0000
		1000	0.7981	0.9444	0.9444	**0.0303**	**0.0303**	1.0000
		5000	0.8056	0.9563	0.9563	0.0542	0.0542	1.0000
	3	500	0.6457	0.8099	0.8303	**0.0052**	**0.0007**	0.2982
		1000	0.6675	0.7963	0.7895	**0.0203**	**0.0336**	0.6896
		5000	0.7319	0.8936	0.8936	**0.0024**	**0.0024**	1.0000
	5	500	0.7217	0.8988	0.9038	**0.0006**	**0.0003**	0.7784
		1000	0.7733	0.8780	0.8780	**0.0248**	**0.0248**	1.0000
		5000	0.7827	0.8973	0.8973	**0.0131**	**0.0131**	1.0000
Structural perturbations	1	500	0.9788	0.9650	0.9956	0.6251	0.1481	0.2895
		1000	0.9888	0.9706	0.9981	0.5227	0.2029	0.3198
		5000	0.9981	1.0000	1.0000	0.3850	0.3850	1.0000
	3	500	0.9836	0.9800	0.9868	0.6323	0.7843	0.3265
		1000	0.9804	0.9840	0.9964	0.8271	0.1088	0.3201
		5000	0.9892	0.9992	0.9992	0.1684	0.1684	1.0000
	5	500	0.9721	0.9571	0.9825	0.5306	0.4884	0.1709
		1000	0.9812	0.9908	0.9958	0.2584	0.1509	0.2200
		5000	0.9975	0.9921	0.9921	0.4761	0.4761	1.0000

Table 16: Statistical difference detection AUCs comparing the uni-model, multi-model, and logistic regression (all using variable ordering information) approaches on tests using the *balance-scale* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.7031	0.8931	0.8931	**0.0074**	**0.0074**	1.0000
		1000	0.7981	0.9444	0.9444	**0.0303**	**0.0303**	1.0000
		5000	0.8056	0.9563	0.9563	0.0542	0.0542	1.0000
	3	500	0.6457	0.8295	0.8303	**0.0008**	**0.0007**	0.4127
		1000	0.6675	0.8043	0.7895	**0.0123**	**0.0336**	0.3189
		5000	0.7319	0.8936	0.8936	**0.0024**	**0.0024**	1.0000
	5	500	0.7217	0.9142	0.9038	**0.0001**	**0.0003**	0.3203
		1000	0.7733	0.8780	0.8408	**0.0248**	0.2020	0.3186
		5000	0.7827	0.8973	0.8973	**0.0131**	**0.0131**	1.0000
Structural perturbations	1	500	0.9788	0.9669	0.9981	0.6727	0.0888	0.2790
		1000	0.9888	0.9706	0.9981	0.5227	0.2029	0.3198
		5000	0.9981	1.0000	1.0000	0.3850	0.3850	1.0000
	3	500	0.9836	0.9800	0.9868	0.6323	0.7843	0.3265
		1000	0.9804	0.9840	0.9964	0.8271	0.1088	0.3201
		5000	0.9892	0.9992	0.9992	0.1684	0.1684	1.0000
	5	500	0.9721	0.9571	0.9825	0.5306	0.4884	0.1709
		1000	0.9812	0.9908	0.9958	0.2584	0.1509	0.2200
		5000	0.9975	0.9921	0.9921	0.4761	0.4761	1.0000

Table 17: Statistical difference detection AUCs comparing the uni-model (using no variable ordering information), multi-model (using no variable ordering information), and logistic regression (using variable ordering information) approaches on tests using the *car* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.5033	0.7058	0.6587	**0.0008**	**0.0282**	0.2110
		1000	0.5744	0.7183	0.7171	**0.0009**	**0.0008**	0.8740
		5000	0.6477	0.7308	0.7221	**0.0409**	0.0604	0.2153
	3	500	0.6296	0.7887	0.7544	$< \mathbf{10^{-4}}$	**0.0026**	0.1901
		1000	0.6595	0.8180	0.7981	$< \mathbf{10^{-4}}$	**0.0001**	0.1675
		5000	0.7136	0.8291	0.8260	**0.0001**	**0.0001**	0.6288
	5	500	0.6149	0.7471	0.7540	**0.0001**	**0.0001**	0.7480
		1000	0.6867	0.7638	0.7567	**0.0151**	**0.0265**	0.2126
		5000	0.7209	0.7887	0.8079	**0.0235**	**0.0018**	0.0673
Structural perturbations	1	500	0.8525	0.9250	0.9229	0.1220	0.1168	0.8424
		1000	0.9096	0.9463	0.9367	0.2765	0.6049	0.7143
		5000	0.9442	0.9804	0.9788	0.0640	0.0696	0.3634
	3	500	0.8952	0.8263	0.8272	**0.0453**	**0.0439**	0.9635
		1000	0.8920	0.8872	0.8739	0.8835	0.6019	0.5784
		5000	0.9572	0.9497	0.9606	0.7599	0.8760	0.5549
	5	500	0.8891	0.8164	0.7774	**0.0395**	**0.0034**	0.1285
		1000	0.9340	0.8881	0.8695	0.1261	0.0593	0.3612
		5000	0.9538	0.9079	0.9043	0.1257	0.1347	0.8486

Table 18: Statistical difference detection AUCs comparing the uni-model, multi-model, and logistic regression (all using variable ordering information) approaches on tests using the *car* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.5033	0.6954	0.6650	**0.0001**	**0.0136**	0.4363
		1000	0.5744	0.7212	0.7271	**0.0008**	**0.0005**	0.2162
		5000	0.6477	0.7188	0.7188	0.1006	0.1006	1.0000
	3	500	0.6296	0.7797	0.7716	$< \mathbf{10^{-4}}$	**0.0004**	0.7681
		1000	0.6595	0.8193	0.8105	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	0.7072
		5000	0.7136	0.8368	0.8368	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	1.0000
	5	500	0.6149	0.7626	0.7767	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	0.5582
		1000	0.6867	0.8038	0.8073	**0.0002**	**0.0001**	0.8155
		5000	0.7209	0.8447	0.8447	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	1.0000
Structural perturbations	1	500	0.8525	0.9308	0.9421	0.0912	**0.0495**	0.1832
		1000	0.9096	0.9521	0.9504	0.1783	0.4317	0.9504
		5000	0.9442	0.9854	0.9854	**0.0207**	**0.0207**	1.0000
	3	500	0.8952	0.8621	0.9164	0.2731	0.5592	0.2214
		1000	0.8920	0.9368	0.9778	0.0853	**0.0105**	0.2053
		5000	0.9572	1.0000	1.0000	**0.0049**	**0.0049**	1.0000
	5	500	0.8891	0.8768	0.9555	0.6474	**0.0012**	**0.0016**
		1000	0.9340	0.9487	0.9820	0.3705	**0.0203**	0.0906
		5000	0.9538	0.9961	0.9961	**0.0101**	**0.0101**	1.0000

Table 19: Statistical difference detection AUCs comparing the uni-model (using no variable ordering information), multi-model (using no variable ordering information), and logistic regression (using variable ordering information) approaches on tests using the *hayes-roth* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.6944	0.7244	0.7256	0.6862	0.6718	0.9212
		1000	0.7375	0.7625	0.7750	0.7222	0.5904	0.1877
		5000	0.6275	0.6713	0.6713	0.5563	0.5563	1.0000
	3	500	0.5889	0.7768	0.7648	**0.0002**	**0.0012**	0.4124
		1000	0.5990	0.7202	0.7210	**0.0311**	**0.0277**	0.9536
		5000	0.5978	0.6925	0.6973	0.0839	0.0634	0.5583
	5	500	0.6766	0.7183	0.7560	0.4247	0.1211	0.1793
		1000	0.7088	0.7371	0.7490	0.5580	0.4394	0.4339
		5000	0.6830	0.7326	0.7331	0.2620	0.2590	0.9563
Structural perturbations	1	500	0.8844	0.8219	0.7763	0.1116	**0.0401**	0.2518
		1000	0.9137	0.8156	0.8081	**0.0070**	**0.0047**	0.4521
		5000	0.9237	0.8425	0.8344	**0.0092**	**0.0058**	0.2232
	3	500	0.8998	0.9103	0.8814	0.6764	0.6295	0.3540
		1000	0.9026	0.9038	0.9387	0.9666	0.2959	0.1712
		5000	0.9131	0.8978	0.8666	0.6553	0.2554	0.2377
	5	500	0.9479	0.9049	0.8546	**0.0471**	**0.0137**	0.0752
		1000	0.9505	0.9288	0.9214	0.1113	0.1106	0.5842
		5000	0.9674	0.9128	0.9136	**0.0248**	**0.0297**	0.8917

Table 20: Statistical difference detection AUCs comparing the uni-model, multi-model, and logistic regression (all using variable ordering information) approaches on tests using the *hayes-roth* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.6944	0.7844	0.7844	0.2031	0.2031	1.0000
		1000	0.7375	0.8194	0.8194	0.2125	0.2125	1.0000
		5000	0.6275	0.8006	0.8006	**0.0109**	**0.0109**	1.0000
	3	500	0.5889	0.8462	0.8442	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	0.6747
		1000	0.5990	0.8422	0.8422	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	1.0000
		5000	0.5978	0.8515	0.8515	$< \mathbf{10^{-4}}$	$< \mathbf{10^{-4}}$	1.0000
	5	500	0.6766	0.8482	0.8388	**0.0002**	**0.0004**	0.3266
		1000	0.7088	0.8591	0.8591	**0.0003**	**0.0003**	1.0000
		5000	0.6830	0.8621	0.8621	**0.0001**	**0.0001**	1.0000
Structural perturbations	1	500	0.8844	0.9137	0.9163	0.1662	0.1434	0.2769
		1000	0.9137	0.9044	0.9044	0.7820	0.7820	1.0000
		5000	0.9237	0.9475	0.9475	**0.0208**	**0.0208**	1.0000
	3	500	0.8998	0.9111	0.9159	0.6443	0.5198	0.2536
		1000	0.9026	0.9131	0.9131	0.5281	0.5281	1.0000
		5000	0.9131	0.9411	0.9411	0.0553	0.0553	1.0000
	5	500	0.9479	0.9405	0.9440	0.6045	0.7906	0.2388
		1000	0.9505	0.9418	0.9418	0.3462	0.3462	1.0000
		5000	0.9674	0.9497	0.9497	0.0985	0.0985	1.0000

Table 21: Statistical difference detection AUCs comparing the uni-model (using no variable ordering information), multi-model (using no variable ordering information), and logistic regression (using variable ordering information) approaches on tests using the *nursery* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.6153	0.7369	0.6234	0.1164	0.9283	0.1347
		1000	0.6247	0.7388	0.6644	0.1718	0.6983	0.2101
		5000	0.6931	0.8641	0.8263	**0.0088**	**0.0499**	0.4076
	3	500	0.5845	0.6503	0.6460	0.2208	0.2978	0.9006
		1000	0.5596	0.6944	0.7179	**0.0089**	**0.0035**	0.4679
		5000	0.5852	0.7774	0.6833	**0.0006**	0.0987	**0.0008**
	5	500	0.6105	0.6479	0.6156	0.4293	0.9226	0.2958
		1000	0.5776	0.6655	0.6149	0.0644	0.4796	0.1172
		5000	0.5755	0.7293	0.7198	**0.0019**	**0.0032**	0.7119
Structural perturbations	1	500	0.7984	0.6528	0.5897	**0.0035**	**0.0035**	0.4192
		1000	0.8469	0.7688	0.7781	0.2244	0.2932	0.8752
		5000	0.8372	0.8678	0.8534	0.6061	0.7410	0.5878
	3	500	0.8148	0.7329	0.7686	**0.0498**	0.2443	0.3103
		1000	0.8527	0.7900	0.8487	0.0987	0.8924	0.0790
		5000	0.8895	0.8967	0.9251	0.8485	0.2596	0.4133
	5	500	0.8326	0.7745	0.8142	0.0855	0.5419	0.1060
		1000	0.8345	0.7622	0.8125	**0.0355**	0.4145	0.0818
		5000	0.8919	0.9351	0.9270	0.0566	0.1333	0.6407

Table 22: Statistical difference detection AUCs comparing the uni-model, multi-model, and logistic regression (all using variable ordering information) approaches on tests using the *nursery* data.

			AUC			p-value		
			λ	Uni	Multi	U v. λ	M v. λ	U v. M
Parametric perturbations	1	500	0.6153	0.7291	0.6291	0.1378	0.8796	0.1995
		1000	0.6247	0.7453	0.6703	0.1497	0.6545	0.2146
		5000	0.6931	0.8641	0.8263	**0.0088**	**0.0499**	0.4076
	3	500	0.5845	0.6552	0.6577	0.1927	0.2214	0.9426
		1000	0.5596	0.7023	0.7364	**0.0062**	**0.0011**	0.3041
		5000	0.5852	0.7896	0.6953	**0.0003**	0.0654	**0.0009**
	5	500	0.6105	0.6413	0.5977	0.5152	0.8015	0.1626
		1000	0.5776	0.6815	0.6115	**0.0278**	0.5199	**0.0322**
		5000	0.5755	0.7310	0.7108	**0.0014**	**0.0055**	0.4102
Structural perturbations	1	500	0.7984	0.6538	0.5916	**0.0041**	**0.0041**	0.4414
		1000	0.8469	0.7647	0.8122	0.1744	0.5994	0.4848
		5000	0.8372	0.8700	0.8553	0.5787	0.7134	0.5565
	3	500	0.8148	0.7220	0.7913	**0.0250**	0.5498	0.0680
		1000	0.8527	0.7937	0.8656	0.1198	0.6607	**0.0402**
		5000	0.8895	0.9028	0.9455	0.7250	0.0597	0.2175
	5	500	0.8326	0.7740	0.8281	0.0825	0.8767	**0.0380**
		1000	0.8345	0.7622	0.8301	**0.0355**	0.8729	**0.0186**
		5000	0.8919	0.9418	0.9418	**0.0242**	**0.0292**	1.0000

5.3 EVALUATION OF CLINICAL SIGNIFICANCE DETECTION

I evaluated the performances of the four test in Section 4.4. These evaluations used the semi-synthetic data described in Section 5.1. Recall that in order to apply a clinical significance test, we must first have a BN model. The BN model used for applying the tests is the one constructed using the multi-model method, as in Section 4.3.4.

As above, the evaluation metric used is the AUC, where AUC's were computed by scoring each node in each of the 20 pairs of groups of data in each test block. The score of a variable X_i used for evaluating computing the AUC for each of the tests in Section 4.4 was defined as:

$$\max_{j=1}^{J_i} \max_{k=1}^{K_i} |P(x_{ik}|\pi_{ij}, Z = 2) - P(x_{ik}|\pi_{ij}, Z = 1)| \tag{5.2}$$

for the difference test,

$$\max_{j=1}^{J_i} \max_{k=1}^{K_i} |\log P(x_{ik}|\pi_{ij}, Z = 2) - \log P(x_{ik}|\pi_{ij}, Z = 1)| \tag{5.3}$$

for the ratio test,

$$\operatorname*{arginf}_{\delta} \max_{j=1}^{J_i} \max_{k=1}^{K_i} P(|P(x_{ik}|\pi_{ij}, Z = 2) - P(x_{ik}|\pi_{ij}, Z = 1)| > \delta) > 1 - \alpha \tag{5.4}$$

the probabilistic difference test, and

$$\operatorname*{arginf}_{\varepsilon} \max_{j=1}^{J_i} \max_{k=1}^{K_i} P(|\log P(x_{ik}|\pi_{ij}, Z = 2) - \log P(x_{ik}|\pi_{ij}, Z = 1)| > \varepsilon) > 1 - \alpha \tag{5.5}$$

for the probabilistic ratio test, with $\alpha = 0.05$ for both probabilistic tests.

The AUCs obtained for all tests therefore correspond to varying the δ and ε thresholds.

5.3.1 Gold standard

To evaluate the detection of clinically significant differences, we must first determine which differences are clinically significant in the generating models of the data. From the experimental setup we know which variables are perturbed in the generating models of each pair of semi-synthetic groups of data. However, perturbation of the generating network alone does not guarantee that the conditional probabilities of a variable were changed sufficiently to constitute a clinical difference.

In order to determine which perturbations of the generating model caused clinically significant differences to appear, we must compare the corresponding conditional probabilities. For the probabilities of variables to which parametric perturbations were applied, the comparison is simple: X_i shares the same parents in both generating BNs, and we can therefore compare each $P(x_{ik}|\pi_{ij})$ in one BN against $P(x_{ik}|\pi_{ij})$ in the second BN. For variables to which a structural perturbation was applied, however, the parent sets in the two BNs, $\Pi_i^{(1)}$ and $\Pi_i^{(2)}$, are different. Let $\Pi_i^* = \Pi_i^{(1)} \cup \Pi_i^{(2)}$. To determine whether X_i has a clinically significant perturbation, I compare each $P(x_{ik}|\pi_{ij}^*)$ between the the two networks, over all instantiations of the union of parent sets.

For the purposes of this evaluation, I defined the threshold for clinical significance as follows: a pair of conditional probabilities (compared as above) is considered to have a clinically significant difference if one is greater than the other by a factor of more than two.

5.3.2 Results

Tables 23, 24, 25, and 26 show the AUC's for clinical difference recovery on the 72 blocks of tests. Each table corresponds to a UCI Repository data source and is structured as follows: The first column indicates the perturbation type (structural or parametric), the second column indicates the number of perturbations, and the third column indicates the number of data points per group used in each test. Since a model is required for the clinical significance tests, the results show AUCs from the four tests both using a multi-model learned without ordering informations and a multi-model learned with ordering information. For each of the two groups of tests I show in bold those AUCs which are best for the group.

When multiple AUCs are tied for best, all are shown in bold.

Overall, of 72 test blocks, when the underlying model was learned without ordering information, the absolute difference test performed best (ties included) in 30 blocks, the absolute ratio test performed best in 38 blocks, the probabilistic difference test performed best in 11 blocks, and the probabilistic ratio test performed best in 13 blocks. When the underlying model was learned with ordering information, the absolute difference test performed best in 38 test blocks, the absolute ratio test performed best in 36 test blocks, and the probabilistic tests each performed best in only 9 test blocks.

The most surprising result that the evaluation reveals is that more often than not, the probabilistic tests perform worse than the non-probabilistic tests. The results above only show that the probabilistic tests have poorer detection at the $\alpha = 0.05$ threshold. It is possible that better AUCs for the probabilistic tests can be obtained by varying not only the ε and δ thresholds, but also the α threshold.

Trends that are common to all tests are that less variables, more data points per group, and ordering information, all yield higher detection AUCs.

Table 23: AUCs for the various clinical significance tests using the *balance-scale* data.

			Without ordering information				With ordering information			
			Diff.	Rat.	P. Diff.	P. Rat.	Diff.	Rat.	P. Diff.	P. Rat.
Parametric perturbations	1	500	0.9992	**1.0000**	0.9975	0.9892	0.9992	**1.0000**	0.9975	0.9892
		1K	**0.9975**	0.9958	0.9900	0.9743	**0.9975**	0.9958	0.9900	0.9743
		5K	**1.0000**	**1.0000**	0.9867	0.9751	**1.0000**	**1.0000**	0.9867	0.9743
	3	500	0.9807	**0.9812**	0.9741	0.9696	0.9807	**0.9812**	0.9741	0.9696
		1K	0.9762	**0.9807**	0.9543	0.9330	0.9762	**0.9807**	0.9538	0.9330
		5K	0.9939	**1.0000**	0.9508	0.9300	0.9939	**1.0000**	0.9508	0.9300
	5	500	**0.9134**	**0.9134**	0.9034	0.9006	**0.9134**	**0.9134**	0.9034	0.9006
		1K	0.9396	**0.9420**	0.9207	0.9147	0.9396	**0.9420**	0.9215	0.9147
		5K	0.9960	**0.9984**	0.9541	0.9424	0.9960	**0.9984**	0.9541	0.9420
Structural perturbations	1	500	**0.9087**	0.9063	0.9081	0.9056	0.9184	0.9159	**0.9203**	0.9178
		1K	0.9688	**0.9750**	**0.9750**	**0.9750**	0.9688	**0.9750**	**0.9750**	**0.9750**
		5K	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	3	500	**0.9804**	**0.9804**	**0.9804**	**0.9804**	**0.9804**	**0.9804**	**0.9804**	**0.9804**
		1K	**0.9706**	**0.9706**	**0.9706**	**0.9706**	**0.9706**	**0.9706**	**0.9706**	**0.9706**
		5K	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	5	500	**0.9746**	**0.9746**	**0.9746**	**0.9746**	**0.9746**	**0.9746**	**0.9746**	**0.9746**
		1K	**0.9831**	**0.9831**	**0.9831**	**0.9831**	**0.9831**	**0.9831**	**0.9831**	**0.9831**
		5K	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Table 24: AUCs for the various clinical significance tests using the *car* data.

			Without ordering information				With ordering information			
			Diff.	Rat.	P. Diff.	P. Rat.	Diff.	Rat.	P. Diff.	P. Rat.
Parametric perturbations	1	500	0.7965	**0.7973**	0.7935	0.7950	**0.8392**	0.8377	0.7746	0.7777
		1K	**0.8954**	0.8923	0.8815	0.8862	**0.9419**	0.9335	0.8635	0.8850
		5K	**0.9946**	0.9923	0.9831	0.9915	**1.0000**	0.9915	0.8169	0.8385
	3	500	**0.8847**	0.8838	0.8650	0.8675	**0.9011**	0.8960	0.8192	0.8224
		1K	0.9809	**0.9812**	0.9777	0.9761	**0.9901**	0.9844	0.8948	0.9059
		5K	0.9866	**0.9904**	0.9751	0.9831	0.9745	**0.9809**	0.8504	0.8565
	5	500	**0.8643**	0.8592	0.8526	0.8526	**0.8604**	0.8563	0.8380	0.8355
		1K	**0.8679**	0.8577	0.8529	0.8549	**0.9350**	0.9241	0.8726	0.8842
		5K	**0.9586**	0.9573	0.9259	0.9221	**0.9695**	0.9627	0.8035	0.8116
Structural perturbations	1	500	0.8635	0.8615	**0.8644**	0.8552	0.9075	**0.9104**	0.8796	0.8621
		1K	0.9342	0.9363	**0.9375**	0.9283	**0.9442**	0.9413	0.9075	0.9013
		5K	**0.9644**	0.9610	0.9569	0.9548	0.9648	**0.9715**	0.8931	0.8535
	3	500	0.8036	0.8101	0.7934	**0.8133**	**0.8809**	0.8779	0.8653	0.8641
		1K	0.8485	**0.8782**	0.8313	0.8526	0.9612	**0.9704**	0.9161	0.9111
		5K	0.9370	**0.9733**	0.9219	0.9192	0.9939	**1.0000**	0.9049	0.8922
	5	500	0.7498	**0.7567**	0.7365	0.7296	0.9054	**0.9085**	0.8962	0.8944
		1K	0.8612	**0.8827**	0.8441	0.8614	0.9874	**0.9877**	0.9562	0.9542
		5K	0.8873	**0.9167**	0.8920	0.8914	0.9953	**0.9992**	0.9167	0.9178

Table 25: AUCs for the various clinical significance tests using the *hayes-roth* data.

			Without ordering information				With ordering information			
			Diff.	Rat.	P. Diff.	P. Rat.	Diff.	Rat.	P. Diff.	P. Rat.
Parametric perturbations	1	500	0.8764	0.8804	0.8723	**0.8818**	0.7242	**0.7473**	0.5693	0.5353
		1K	0.8621	0.8662	0.8662	**0.8743**	0.7351	**0.7527**	0.5530	0.5951
		5K	0.9606	0.9755	0.9606	**0.9783**	**0.9443**	**0.9443**	0.6318	0.6223
	3	500	**0.7582**	0.7550	0.7574	0.7560	**0.7046**	0.7001	0.5826	0.5201
		1K	**0.7646**	0.7567	0.7597	0.7423	**0.6731**	0.6277	0.5675	0.5308
		5K	**0.8924**	0.8735	0.8452	0.8418	**0.8395**	0.8167	0.6094	0.5434
	5	500	0.7743	0.7671	0.7730	**0.7823**	**0.6990**	0.6944	0.6608	0.5931
		1K	0.7814	0.7806	**0.7928**	0.7617	**0.7053**	0.6915	0.6566	0.6259
		5K	**0.8869**	0.8840	0.8558	0.8567	**0.8806**	0.8436	0.6818	0.6028
Structural perturbations	1	500	0.5014	0.4943	0.4979	**0.5106**	0.8717	**0.8753**	0.5443	0.5514
		1K	**0.6531**	0.6304	0.6347	0.6283	**0.8887**	0.8738	0.5138	0.5656
		5K	0.8731	**0.8866**	0.8200	0.8859	**0.9787**	0.9724	0.5507	0.5734
	3	500	0.8212	0.8258	0.8315	**0.8563**	0.9632	**0.9682**	0.8408	0.8727
		1K	0.8743	0.8739	0.8396	**0.9074**	**0.9804**	0.9779	0.7848	0.8191
		5K	**0.8131**	0.8111	0.7561	0.7979	**0.9979**	0.9975	0.7400	0.7916
	5	500	**0.8847**	0.8746	0.8360	0.8045	**0.9223**	0.9199	0.8466	0.8539
		1K	0.9585	0.9642	0.9328	**0.9662**	**0.9269**	0.9261	0.8368	0.8380
		5K	**0.9340**	0.9263	0.8788	0.8893	**0.9477**	0.9469	0.8349	0.8450

Table 26: AUCs for the various clinical significance tests using the *nursery* data.

			Without ordering information				With ordering information			
			Diff.	Rat.	P. Diff.	P. Rat.	Diff.	Rat.	P. Diff.	P. Rat.
Parametric perturbations	1	500	**0.8670**	0.8663	0.8554	0.8590	**0.8568**	0.8539	0.8234	0.8328
		1K	**0.8714**	0.8670	0.8568	0.8656	**0.8641**	0.8597	0.8307	0.8430
		5K	**0.9033**	0.8997	0.8517	0.8946	**0.9033**	0.8997	0.8517	0.8939
	3	500	0.7476	**0.7522**	0.7483	0.7486	0.7015	**0.7110**	0.6806	0.6904
		1K	0.7679	**0.7728**	0.7601	**0.7728**	0.7519	**0.7594**	0.7300	0.7496
		5K	0.7766	**0.7945**	0.7223	0.7926	0.7681	**0.7890**	0.6952	0.7857
	5	500	0.7471	**0.7528**	0.7316	0.7517	0.7053	**0.7119**	0.6589	0.6937
		1K	0.7272	**0.7281**	0.7007	0.7277	0.7318	**0.7322**	0.6781	0.7172
		5K	0.8082	**0.8100**	0.7031	0.7908	0.7812	**0.7844**	0.6639	0.7659
Structural perturbations	1	500	**0.5404**	**0.5404**	0.5397	**0.5404**	**0.5768**	0.5765	0.5680	0.5700
		1K	0.6067	0.6074	0.6028	**0.6080**	0.6702	**0.6734**	0.6496	0.6626
		5K	0.8395	**0.8571**	0.8343	0.8542	0.8415	**0.8578**	0.8326	0.8535
	3	500	0.6710	0.6740	0.6631	**0.6743**	0.7110	**0.7126**	0.7036	**0.7126**
		1K	0.7285	**0.7336**	0.7175	0.7317	0.7588	**0.7637**	0.7447	0.7551
		5K	0.9012	**0.9102**	0.8662	0.9064	0.9269	**0.9304**	0.9026	0.9181
	5	500	0.7149	**0.7172**	0.7076	0.7163	0.7273	**0.7285**	0.7193	0.7253
		1K	0.7897	**0.7904**	0.7750	0.7871	0.8322	**0.8322**	0.8142	0.8290
		5K	0.9091	**0.9107**	0.8690	0.9078	0.9152	**0.9170**	0.8734	0.9053

6.0 EXPLAINING DIFFERENCES OF DISTRIBUTIONS

As discussed in the Chapter 1, a good answer to the question of what is different between two groups of data consists not only of listing the differences, but of using a model that captures the differences and similarities, and showing how differences and similarities in the model combine to explain the observed data. Chapter 4 developed two approaches (uni-model and multi-model) to detecting statistical differences and incorporating those differences in a BN model. In this chapter, I present methods that use such models to explain differences.

Particularly, the differences that I focus on explaining are the differences in the marginal distribution of a variable X_i. The difference in the marginal distribution of a variable is one of the easiest differences to identify directly from data, but without further explanation, the presence of such a difference can become puzzling. I call the explanation approaches developed here "explanation by traversal" because of the relationship between the explanation process and the BN structure of the model used to explain a marginal difference.

The general process of an explanation by traversal requires as input a BN model produced using either the uni-model approach or the model construction process from the multi-model approach. The explanation process starts by examining a difference in the marginal probability of X_i taking a value x_{ik}, as computed according to the model. The next step is to explain the local factors that contribute to the observed difference in terms of the probabilistic relationship between X_i, the parent set Π_i, and the group indicator Z. Subsequent steps consist of examining the parents of X_i, relating the differences in their distributions individually to the difference of interest, and if necessary, repeating the process. The process of explanation starts at X_i and traverses upwards in the ancestry of X_i. It can also be seen as tracing the BN inference process of computing the difference in probabilities of x_{ik}.

A broader view of how this fits with the methods in Chapter 4 is that the tests of

statistical significance determine which parametric differences appear in the BN model. The BN model structure determines which variables are included in the explanation for the observed difference in each variable. The tests of clinical significance determine which factors that contribute to a difference are important enough to be included in the explanation of the difference.

The parameter differences found by statistical difference testing are used to determine which variables have Z as a parent, and hence, which variables in the model have conditional distributions that are different across the groups. With respect to the model, such differences are *elementary* in the sense of having no explanation in terms of any other information present in the model. These elementary differences also determine which variables in the model can in principle have different marginal distributions across the groups, namely, only variables that have Z as an ancestor (but not necessarily a parent) can have differences in their marginal distributions. Therefore, the most detailed explanation possible of the difference in the marginal probability of x_{ik}, as computed according to the model, is an account of how each elementary difference in each ancestor of X_i that has Z as a parent contributes to the difference in the marginal probability of x_{ik}. When X_i has multiple ancestors that depend on Z, a full account of this form becomes quite unwieldy in the amount of detail that it includes. Moreover, it is in practice often the case that only some of the many elementary parameter differences across the two models are driving most of the observed difference in the marginal probability of x_{ik}. For this reason, in the explanation methods developed below, I filter contributions using clinical significance test, to focus the explanation only on clinically significant contributions.

The hypothesis explored by developing such explanation methods is that *by traversing the structure of a BN generated using the uni-model or multi-model approach, showing how the elementary differences in the model contribute to the calculation of the difference in the marginal probability of x_{ik}, and focusing only on clinically significant elements of the calculation, we can provide useful insights into the reason for the marginal difference in the probability ofx_{ik}.*

To evaluate this hypothesis, I present case studies of the explanation methods developed here to real clinical data. This chapter describes three approaches to difference explanation

by traversal that I have developed. Ordered from the simplest to the most complex, they are: CPR (Comparison of Probabilistic Relationships), EDAPD (Explanation of Differences Across a Pair of Datasets), and DECC (Difference Explanation by Condition Carrying). In CPR I introduce the concept of explanation of a difference by local decomposition based on a BN model; in EDAPD I introduce the application of recursion in order to generate deeper explanations, but introduce constraints on the network structure to facilitate recursion; finally, after discussing potential alternate approaches, I introduce DECC, which does not require these structural constraints.

6.1 COMPARISON OF PROBABILISTIC RELATIONSHIPS

CPR is an approach to difference explanation that aims to explain the difference in the distribution of a variable of interest X_i using the local BN structure. It presents an explanation in terms of contributions from differences in the the parameters of X_i between the two groups and contributions from the differences in the distributions X_i's parents in the BN (Sverchkov et al., 2012).

Here we are interested in explaining the difference in X_i's distribution between $\mathcal{D}_1$ and $\mathcal{D}_2$. The approach taken in CPR is to look for large terms that contribute to the mathematical difference between $P(x_{ik}|Z=1)$ and $P(x_{ik}|Z=2)$ for each value x_{ik} that X_i takes. For simplicity and consistency, I try to always represent positive differences (and when the discussion moves on to quotients, terms will be arranged to keep quotients above 1). To achieve this, let $z_1 := \text{argmax}_z P(x_{ij}|z)$, making $P(x_{ik}|z_1)$ the larger probability and $P(x_{ik}|\neg z_1)$ the smaller probability. The difference $P(x_{ik}|z_1) - P(x_{ik}|\neg z_1)$ then becomes the quantity that we want to explain.

When the node X_i has parents other than Z in the BN, we explain the difference in the distribution of X_i between the two groups in terms of its parents. Noting the equality

$$P(x_{ik}|z) = \sum_{j=1}^{J_i} P(x_{ik}, \pi_{ij}|z) \tag{6.1}$$

it is clear that the difference of interest decomposes as follows:

$$P(x_{ik}|z_1) - P(x_{ik}|1 - z_1) = \sum_{j=1}^{J_i} \left[P(x_{ik}, \pi_{ij}|z_1) - P(x_{ik}, \pi_{ij}|\neg z_1) \right]. \tag{6.2}$$

Let us refer to the difference $P(x_{ik}|z_1) - P(x_{ik}|\neg z_1)$ as the *marginal difference term*, and refer to each difference of the form $P(x_{ik}, \pi_{ij}|z_1) - P(x_{ik}, \pi_{ij}|\neg z_1)$ as a *joint difference term*. Equation (6.2) shows that for each assignment of the parents π_{ij}, the joint difference term contributes either towards or against the marginal difference term, depending on its sign. Since the number of parent configurations J_i grows exponentially with the number of parents $|\Pi_i|$, the number of joint terms can be quite large. In order to keep the explanation at a size that is manageable for a user to view and understand, CPR filters the explanations, and to further facilitate the presentation, the joint terms that pass the filter are grouped by sign and sorted by magnitude in descending order. The filter used in (Sverchkov et al., 2012) is the absolute difference test (4.29) described in Section 4.4 on the event x_{ik}, π_{ij}. The threshold δ described there indirectly controls the number of terms in the sum that are displayed: a smaller δ displays more terms ($\delta = 0$ displays all terms) and a larger δ displays less terms ($\delta > 1$ displays no terms). Other filtering rules are also possible, for example, one could only show enough terms to account for a certain proportion of the marginal difference term, or one could filter by any other test discussed in Section 4.4. Conceptually, there is an appeal to using tests based on a difference of the probabilities, since the sum of the differences is the observed marginal difference.

Figure 8 is an algorithmic summary of the process of generating the explanation, where:

1. Report$(x_{ik}|Z)$ is a report of the marginal probability of x_{ij} in each group. In (Sverchkov et al., 2012) this included listing the probabilities $P(x_{ik}|Z = 1)$, $P(x_{ik}|X = 2)$ and the difference $P(x_{ik}|Z = 2) - P(x_{ik}|X = 1)$. A more detailed report might also include Bayesian credibility intervals for this difference, and similar statistics for the ratio of the two probabilities.
2. ClinicalSignificanceTest$((x_{ik}, \pi_{ij})$ is the filtering test discussed above.
3. Report$(x_{ik}, \pi_{ij}|Z)$ is a report of the joint probability of x_{ik}, π_{ij} in each group, discussed in detail below.

```
1: procedure CPR(i,k)
2:     Print Report(x_ik|Z)
3:     PositiveReportList ← ∅
4:     NegativeReportList ← ∅
5:     Let z_1 := argmax_z P(x_ik|z)
6:     for j ∈ {1, ..., J_i} do
7:         if ClinicalSignificanceTest(x_ik, π_ij) then
8:             Report ← Report(x_ik, π_ij|Z)
9:             if P(x_ik, π_ij|z_1) ≥ P(x_ik, π_ij|¬z_1) then
10:                Append(PositiveReportList, Report)
11:            else
12:                Append(NegativeReportList, Report)
13:    Sort(PositiveReportList)
14:    Sort(NegativeReportList)
15:    Print(PositiveReportList)
16:    Print(NegativeReportList)
```

Figure 8: Main CPR procedure.

The explanation does not stop at listing these filtered joint difference terms. The joint difference terms are one way to measure how the joint probabilities $x_{ik}|\pi_{ij}$ differ across the groups. Another measure that we can consider is the ratio of these probabilities, the *joint ratio term*.[1] The joint ratio terms can be further broken down into contributions from different factors. Each joint probability $P(x_{ik}, \pi_{ij}|z)$ has a natural decomposition that is provided by the BN:

$$P(x_{ik}, \pi_{ij}|z) = P(x_{ik}|\pi_{ij}, z)P(\pi_{ij}|z) \ . \tag{6.4}$$

Here, the *conditional probability term* $P(x_{ik}|\pi_{ij}, z)$ is a conditional probability of an assignment of X_i given its parents. When Z is not a parent of X_i, the fact that X_i cannot be an ancestor of Z implies that conditioning on X_i's parents and not conditioning on any descendants of X_i d-separates X_i from Z, making the probability equal to $P(x_{ik}|\pi_{ij})$, meaning that even though the conditional probability term is expressed with conditioning on z, it corresponds to a network parameter both when Z is and when Z isn't a parent of X_i. The term $P(\pi_{ij}|z)$ is the joint probability of the assignment of the parents within one of the groups.

Let $z_2 := \text{argmax}_z \, P(x_{ik}, \pi_{ij}|z)$. The decomposition in (6.4) leads to the decomposition of the joint ratio term as follows:

$$\frac{P(x_{ik}, \pi_{ij}|z_2)}{P(x_{ik}, \pi_{ij}|\neg z_2)} = \frac{P(x_{ik}|\pi_{ij}, z_2)}{P(x_{ik}|\pi_{ij}, \neg z_2)} \frac{P(\pi_{ij}|z_2)}{P(\pi_{ij}|\neg z_2)} \ . \tag{6.5}$$

I refer to the term $\frac{P(x_{ik}|\pi_{ij}, z_2)}{P(x_{ik}|\pi_{ij}, \neg z_2)}$ as the *conditional ratio term*. Consider the right-hand side of Equation (6.5): testing whether a ratio is greater or less than 1 shows whether it contributes towards or against the ratio $\frac{P(x_{ik}, \pi_{ij}|z_2)}{P(x_{ik}, \pi_{ij}|\neg z_2)}$. Moreover, we obtain a measure of the magnitude of the contribution as a multiplicative factor by looking at the value of the ratio for terms

[1] While in general, knowing the ratio between two quantities is insufficient to determine the difference between them (or vice versa), with respect to a fixed joint probability in group z, a linear relationship between the joint difference term and the joint ratio term holds:

$$(P(x_{ik}, \pi_{ij}|\neg z) - P(x_{ik}, \pi_{ij}|z)) \frac{1}{P(x_{ik}, \pi_{ij}|z)} = \frac{P(x_{ik}, \pi_{ij}|\neg z)}{P(x_{ik}, \pi_{ij}|z)} - 1 \ . \tag{6.3}$$

Due to this relationship, factors that contribute to a increasing or decreasing the joint ratio term also contribute to increasing or decreasing, respectively, the joint difference term.

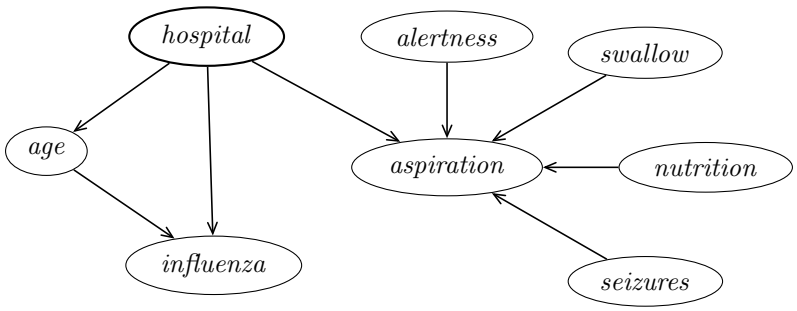

Figure 9: A fragment of the BN learned from the CEHC-PORT data by the greedy thick-thinning algorithm.

that contribute towards $\frac{P(x_{ik},\pi_{ij}|z_2)}{P(x_{ik},\pi_{ij}|\neg z_2)}$ and at the inverse of the value for terms that contribute against the ratio.

Representing the variables in Π_i as $Y_1, \ldots, Y_m$ and their respective assignments in π_{ij} by $y_{1j}, \ldots y_{mj}$, the term $P(\pi_{ij}|z)$ can be decomposed into the product

$$P(\pi_{ij}|z) = \prod_{t=1}^{m} P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, z). \tag{6.6}$$

Using this decomposition, we can arrive at a decomposition of the ratio as a product of *parent ratio terms*:

$$\frac{P(\pi_{ij}|z_2)}{P(\pi_{ij}|\neg z_2)} = \prod_{t=1}^{m} \frac{P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, z_2)}{P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \neg z_2)}. \tag{6.7}$$

The parent ratio terms are organized into contributions towards and against the joint ratio term for presentation purposes.

6.1.1 Clinical data case study

CPR was applied to the CEHC-PORT data described in section 2.4. To demonstrate CPR, data from two of the five medical institutions that appear in the dataset were selected as the pair of groups to compare. I will refer to these institutions as *hospital A* and *hospital B*, or simply as values of the *hospital* variable, A and B. The indicator variable Z corresponds to the *hospital* variable.

The uni-model approach was used. To model the data a single BN was learned using the greedy-thick-thinning BN-learning algorithm (Heckerman, 1999) with the K2 score and the constraint of at most five parents per node and with constraints over the variable ordering. The order of the variables in the network was, ordered from ancestors to descendants, constrained as follows: the *hospital* variable was first, followed by demographic variables such as *age* and *sex*, followed by variables that describe the patient's history and state at admission such as *smoke* (whether the patient smokes) and *influenza* (whether the patient had influenza within six weeks prior to presentation), followed by other variables which represent findings such as test results and other information about the patient's state while in the hospital, and outcome variables such as *dead30* and *dead90* (whether the patient has died within 30 or 90 days after presentation) were last. While the order constraints have a loosely causal and temporal justification, the resultant network is not guaranteed to be causal, and the results must be interpreted probabilistically rather than causally.

The absolute-difference-test threshold selected to use for the purposes of these evaluations was $\delta = 0.01$, as it provided an informative yet manageable level of detail. In a practical application, the threshold would be selected based on the user's (e.g., clinical researcher's) goals and preference for level of detail. In the learned BN, among the 165 variables. To illustrate instances where contributions to a difference come both from conditional probability terms and parent probability terms, I concentrated on examining the variables that were children of the *hospital* variable. There were 15 such variables, each of which values for which the marginal differences exceeded the threshold. In a general exploratory analysis of a dataset, however it may be of interest to also focus on other variables. A clinical significance test may be applied on each variable in the dataset to search for differences of interest

that may require explanations, and CPR can be applied to each value assignment of each of those variables. In the current case study, I selected two of the 15 children of the 'hospital' variable, *influenza* and *aspiration* (aspiration event), to illustrate the features of CPR.

For the *influenza* variable, the marginal probability for the value $influenza = yes$, $P(influenza = yes|hospital)$, is 0.130 for *hospital A* and 0.326 for *hospital B*, yielding a difference of 0.196. In the BN, the variable has only one parent besides *hospital*, namely, *age*. The additive terms that contribute to this difference take the form

$$P(influenza = yes, age|hospital = B) - P(influenza = yes, age|hospital = A) \ . \quad (6.8)$$

The values of *age* (as discretized into ranges) corresponding to terms that have a positive difference that exceeds the 0.01 threshold are: 30-44 years old, 0.082; 18-29 years old, 0.056; 75-90 years old, 0.037; 60-74 years old, 0.011. No terms exceeded the threshold and contributed negatively to the difference.

Proceeding to the second level of analysis for the first of these terms, we compute the ratio

$$\frac{P(influenza = yes, age = \text{30-44}|hospital = B)}{P(influenza = yes, age = \text{30-44}|hospital = A)} = 3.307 \ , \quad (6.9)$$

which is further decomposed into the conditional part

$$\frac{P(influenza = yes|age = \text{30-44}, hospital = B)}{P(influenza = yes|age = \text{30-44}, hospital = A)} = 2.713 \quad (6.10)$$

and the parent part

$$\frac{P(age = \text{30-44}|hospital = B)}{P(age = \text{30-44}|hospital = A)} = 1.415 \ . \quad (6.11)$$

Thus, both parts contribute to the joint ratio term, with the conditional part contributing more, meaning that the difference for the subgroup of patients between 30 and 44 years of age is mostly explained by a higher proportion of 30-44 year-olds who had influenza recently, but the fact that there are proportionally more patients that are 30-44 also contributed to the difference. Similar numbers are observed for the terms corresponding to an *age* value of 18-29 and 75-90, with the notable exception that the parent part of the 75-90 ratio contributes slightly against the additive term's ratio:

$$\frac{P(age = \text{75-90}|hospital = B)}{P(age = \text{75-90}|hospital = A)} = 0.987 \ . \quad (6.12)$$

These results show that the higher proportion of patients with influenza at *hospital B* is explained by the observation that *hospital B* patients who were 18-29, 30-44, and 75-90 had more influenza recently than *hospital A* patients, and additionally, there were proportionally more patients with *age* 18-29 and 30-44 at *hospital B* than *hospital A*.

The analysis for the *aspiration* variable is both more interesting and more complex (see Figure 9). The marginal distribution difference is 0.083 for $aspiration = yes$, with *hospital A* seeing proportionally more aspiration events. The *aspiration* variable has *alertness* (patient alertness level), *swallow* (presence of swallowing disorders), *nutrition* (malnutrition or poor nutritional status), and *seizures* (seizures) as parents. Only two additive terms exceed the difference threshold of $\delta = 0.01$, both correspond to absence of *swallow*, *nutrition* and *seizures*, and both contribute positively to the marginal difference. One term corresponds to $alertness = alert$ (patient is alert) and the other to $alertness = lethargic$ (patient is lethargic). The decomposed ratios of both terms show that the contribution of the conditional term (with a factor between 8 and 9) outweighs the contribution of the parent terms (all with factors close to 1).

These results show that the higher proportion of aspiration events at *hospital A* is primarily explained by the observation that among patients without malnutrition, swallowing disorders, or seizures, who are either alert or lethargic (but not unconscious or comatose), more experience aspiration events at *hospital A* than at *hospital B*. These results could have practical implications to a health official wanting to reduce the rate of aspiration events at *hospital A*. The analysis suggests that while alertness, malnutrition, swallowing disorders, and seizures are predictive of aspiration, the difference in aspiration pneumonia across the hospitals is not due to a difference in the prevalence of these factors between hospitals. It appears that the difference in the prevalence of aspiration pneumonia is either due to a difference in the general care within the hospitals, or due to other factors that are not recorded in the data (such as environmental factors). If the cause is the former, there may be an opportunity to improve the general care of patients without these predisposing factors to prevent aspiration pneumonia. Clearly, if the analysis had shown that a difference in the presence of predisposing factors accounts for the observed difference, it would suggest different corrective actions to explore further.

6.2 EXPLANATION OF DIFFECENCES ACROSS A PAIR OF DATASETS

With EDAPD I aimed to generate deeper explanations that trace influences (inferences) to the underlying parameter differences in the model, in contrast to CPR's explanations that are more local, terminating at the parents of the node of interest. Conceptually, once we explain the difference in the marginal distribution of a variable X_i as having contributions from the differences in the conditional distribution of X_i given its parents Π_i and contributions from the differences in the distributions of the parents Π_i, the next logical step is to explain the differences in the distributions of the parents. In CPR, however, once we trace the difference to the contribution from the joint probability term of the parents Π_i of the node X_i having a particular state of the parents π_{ij}, the potential for conditional dependence between the parents precludes us from breaking up the contribution into independent contributions from individual parents. Mathematically, this is because in general

$$P(\pi_{ij}|z) = \prod_{t=1}^{m} P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, z) \neq \prod_{t=1}^{m} P(y_{tj}|z) \ . \tag{6.13}$$

6.2.1 Almost singly connected networks

There is, however, a class of BN structures for which the inequality in (6.13) becomes an equality. I introduce the term "Almost Singly-Connected Network" (ASCN) to describe a specific class of networks that can be described in terms of SCNs: We define an ASCN as a directed acyclic graph with a special node Z such that Z has no parents, and if Z and the links to its children are removed from the DAG, the resulting network is singly-connected. Figure 10 shows and example of an ASCN and the corresponding SCN that is obtained by removing Z. Figure 11 shows a greedy algorithm for learning an ASCN from data $\mathcal{D}$ over an ordered list of variables $\mathbf{X}$ with respect to a group indicator variable Z.

The algorithm builds the network up starting from a DAG the sole node of which is Z, adding X_i's one-by-one in order. The algorithm therefore requires the variables to be ordered. In addition to the DAG, the algorithm keeps track of the set of connected components of all the variables excluding Z, which is initially empty.

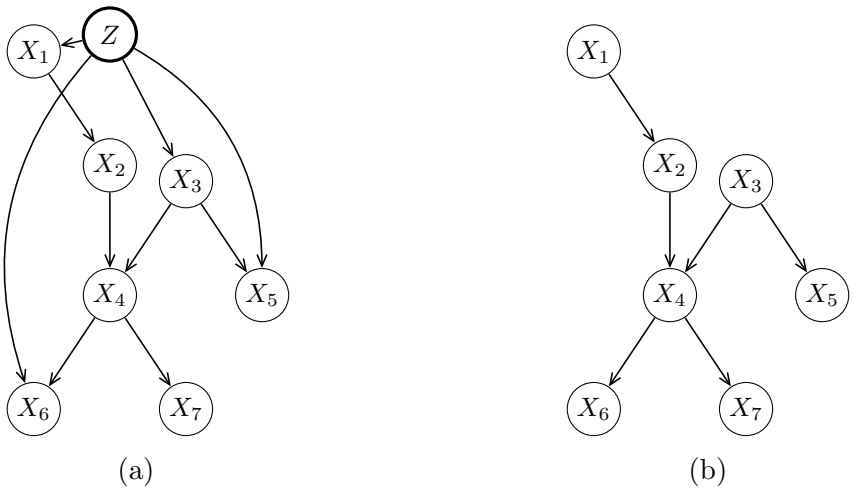

Figure 10: An example of (a) an ASCN, and (b) the SCN that is obtained from it by removing Z.

For each variable X_i, the algorithm iterates over all possible ways to select one or less variable from each of the connected components collected thus far. The algorithm then considers these variables as potential parents of X_i, and computes the score associated with X_i having these variables as parents, and the score associated with X_i having these variables and Z as parents. Having computed the scores for all possible parent-set selections, the highest-scoring parent-set is selected, and the corresponding arcs are added to the network. The set of connected components is updated: the connected components that had a variable in the parent set are merged, and X_i is added to the resulting component. Once the algorithm completes iteration over all X_i's, the resulting DAG is returned as the learned ASCN.

While the search for the potential parents of X_i at each step is exhaustive, the algorithm is greedy in nature and does not guarantee optimality since once the parents for a node are chosen, the set remains fixed for the remainder of the search. This algorithm has a worst-case exponential time complexity because of the exhaustive search over subsets of the set of connected components. In practice, it appears that the number of connected components is typically small. I have used the algorithm to learn an ASCN of over a hundred variables to

```
1: function LearnASCN(𝒟, X, Z)
2:     DAG ← {Z}
3:     ConnectedComponents ← ∅                          ▷ A set of sets of variables.
4:     for i ∈ {1, ..., n} do
5:         AddNode(X_i, DAG)
6:         s ← −∞                                       ▷ Score of candidate parent set.
7:         Parents                                      ▷ Best candidate parent set.
8:         for all Parents'': a selection of one or less variables from each member of
       ConnectedComponents do
9:             for Parents' ∈ {Parents'', Parents'' ∪ {Z}} do
10:                s' ← Score(𝒟, X_i|Parents')
11:                if s' > s then
12:                    Parents ← Parents'
13:                    s ← s'
14:            end for
15:        end for
16:        NewComponent ← {X_i}
17:        for all X_r ∈ Parents do
18:            Let C ∈ ConnectedComponents such that X_r ∈ C
19:            NewComponent ← NewComponent ∪ C
20:            ConnectedComponents ← ConnectedComponents\{C}
21:            AddArc(X_r → X_i, DAG)
22:        end for
23:        ConnectedComponents ← ConnectedComponents ∪ NewComponent
24:    end for
25:    return DAG
```

Figure 11: ASCN structure learner.

learn the BN model in the case study in Section 6.1.1, demonstrating feasibility for practical applications. Other, more greedy and less time complex algorithms are also possible.

6.2.2 Explanation with recursion

To a first approximation, EDAPD is CPR with the addition of recursion for explaining the differences in the distributions of the parent terms, under the assumption that the BN structure is an ASCN. EDAPD is given a variable of interest X_i, the variable whose difference in marginal distribution across the two groups we seek to explain. For each possible value x_{ik} of X_i, EDAPD reports the variable, including any relevant statistics such as point estimates of the probabilities, the expected difference between the probabilities, credibility intervals for the difference, etc.

Next it decomposes the difference into its additive joint components defined by all possible configurations of parents as in (6.2). Typically, a few of the many joint difference terms in the sum dominate the difference; significant terms are filtered using a test from Section 4.4 on the event x_{ik}, π_{ij}. Difference tests are perhaps the most appropriate for this stage since it is the sum of the differences across the groups which adds up to the difference observed in x_{ik}. Again, for each significant joint difference term, the term's expected value and credibility interval are reported.

For each significant joint difference term, EDAPD proceeds by providing a multiplicative decomposition corresponding joint ratio term. Since the network structure is an ASCN, we can express the multiplicative decomposition as follows:

$$\frac{P(x_{ik}, \pi_{ij}|z_2)}{P(x_{ik}, \pi_{ij}|\neg z_2)} = \frac{P(x_{ik}|\pi_{ij}, z_2)}{(x_{ik}|\pi_{ij}, \neg z_2)} \prod_{t=1}^{m} \frac{P(y_{tj}|z_2)}{P(y_{tj}|\neg z_2)} \quad . \tag{6.14}$$

The ASCN assumption guarantees independence between the parents of X_i: since the parents are connected via X_i, they must have no common ancestors other than Z and they cannot be ancestors to each other. Without the ASCN assumption, only a decomposition analogous to equation (6.7) would be valid in general.

The conditional ratio terms and parent ratio terms on the right side of (6.14) can be filtered to reduce the size of the explanation and bring out the most significant terms contributing to the ratio. However, each term has a qualitatively different meaning. The first

```
procedure EDAPD(x_ik)
    REPORT(x_ik|Z)
    Let z_1 := argmax_z P(x_ik|z)
    for j ∈ {1, ..., J_i} do
        if CLINICALSIGNIFICANCETEST(x_ik,π_ij) then
            REPORT(x_ik, π_ij|Z)
            REPORT(x_ik|π_ij, Z)
            for t ∈ {1, ..., m} do
                REPORT(y_tj|Z)
                if CLINICALSIGNIFICANCETEST(y_tj) then
                    EDAPD(y_tj)
```

Figure 12: The recursive EDAPD analysis procedure.

term is a network parameter, and its contribution is considered to be an elementary cause with respect to the underlying model. The rest of the terms are individual contributions from variables other than X_i. For this reason it is helpful to list all the multiplicative terms. It is clearer to explicitly state that while X_i is dependent on the value of Y_t, Y_t has similar distributions across the two groups, rather than (possibly confusingly) omit Y_t from the explanation altogether.

While it is important to list all the multiplicative terms, only those that show significant differences need to be explained. We can use any of the clinical significance tests listed in Section 4.4 on the event y_{tj} to decide whether to generate a further, deeper explanation for each parent ratio term. Since these are terms in a multiplicative decomposition, ratio tests are more appropriate. Once we decide that the marginal difference in a variable Y_t needs to be explained, we repeat the process above, treating y_{tj} as we did x_{ik}. Figure 12 is an algorithmic summary of the explanation procedure. For simplicity I omit explanation sorting and grouping from the pseudocode.

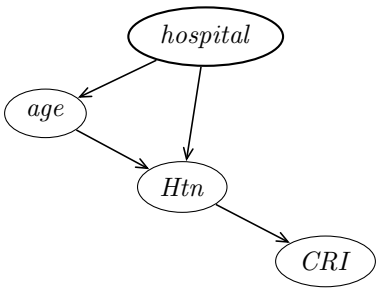

Figure 13: A fragment of the ASCN learned from the CEHC-PORT data.

6.2.3 Clinical data case study

EDAPD was applied to the CEHC-PORT clinical data described in Section 2.4. The uni-model approach was used and a single network structure to capture the data was learned using a greedy ASCN learner and the K2 score. The order of the variables in the structure was the same order used in the case study in Section 6.1.1.

The clinical significance tests applied were the probabilistic absolute difference test with $\delta = 0.01$ and $\alpha = 0.05$ for filtering the additive x_{ik}, π_{ij} terms, and the probabilistic absolute log-ratio test with $\varepsilon = 1.01$ and $\alpha = 0.05$ for filtering marginal parent terms y_{tj}. These numbers were picked because they were expected to lead us to find likely and notable differences while maintaining a reasonable level of detail. In practice, the thresholds can be specified based on the user's needs. A sample of 1000 BN parameterizations were generated by random sampling of the underlying posterior Dirichlet distributions and used to estimate the probabilities in probabilistic clinical significance tests (see Section 4.4).

Of the 165 variables, EDAPD found 57 that showed significant differences at the marginal level. That is, there were 57 variables X_i that had some state x_{ik} which passed the probabilistic absolute difference test. I selected one variable which indicates chronic renal insufficiency (CRI) as the variable to examine in this case study, since it is both clinically interesting and since its analysis illustrates recursive analysis performed by EDAPD. Figure 13 shows the fragment of the learned ASCN that is relevant to the analysis of CRI. Note that the nodes shown in the figure have additional child nodes that are not shown here for brevity.

The probability for the value $CRI = yes$, $P(CRI = yes|hospital)$, is 0.081 for *hospital A* and 0.052 for *hospital B*, yielding a difference of 0.028, with a 95% credibility interval of $(0.017, 0.042)$.

The only parent of the *CRI* variable in the network is the hypertension (*Htn*) variable. The only joint term contributing to the difference is the $P(CRI = yes, Htn = yes|hospital)$ term, which evaluates to 0.0651 for *hospital A* and 0.032 for *hospital B*, yielding a difference of 0.033, with a 95% C.I. of $(0.021, 0.048)$. Thus, the presence of hypertension is accounting for most of the difference in the presence of chronic renal disease between the sites. This translates to a ratio of $\frac{0.065}{0.032} = 2.043$ with a C.I. of $(1.653, 2.460)$.

Since the *hospital* variable is not a direct parent of *CRI*, according to the model, there is no conditional component that contributes to the ratio, and the full contribution to the ratio comes from the parent component $P(Htn = yes|hospital)$. This latter probability is 0.378 for *hospital A* and 0.185 for *hospital B*, yielding a quotient contribution of 2.043 $(1.653, 2.460)$ and a difference of 0.193 $(0.130, 0.254)$. At this point, EDAPD recursively focuses on explaining this difference in the hypertension rate between the two sites.

The *Htn* variable has both the *hospital* and the *age* variable as parents in the network. The joint $P(Htn = yes, age|hospital)$ terms for all *age* values that fall between ages 18 and 74 contribute significantly to the *Htn* difference. The difference contributions, $P(Htn = yes, age|hospitalA) - P(Htn = yes, age|hospitalB)$ were significant for the four age groups that cover the ages 18-74 out of the six possible age groups. Decomposition of the quotients $P(Htn = yes, age|hospitalA)/P(Htn = yes, age|hospitalB)$ revealed that the conditional and parent components contribute about equally for the age group of 30-44, while the conditional component dominates the quotient for the other age groups.

To summarize, EDAPD traces the difference in chronic renal insufficiency between the institutions to differences in hypertension between the institutions. In turn, there are two reasons that hypertension differs between the two hospitals. One is that there are higher rates of hypertension at *hospital A* for four of six different age groups, 18-29, 30-44, 45-59, and 60-74 (the other age groups cover ages 75 and up). Another reason is that the proportion of patients who are 30-44 is higher at *hospital A*.

6.3 RECURSION WHEN PARENTS ARE DEPENDENT

Recursion in EDAPD is straightforward because the parents of a node are guaranteed to be independent when conditioned on Z. The ability to express the joint distribution of the parents of X_i as a product of marginal distribution of individual parents ensures consistency across the various stages of the explanation. In general BN's, however, there is no such guarantee of independence, and we are therefore tasked with finding a different approach to ensuring consistency across explanation stages if we are to allow for recursion.

Dependencies between parents are manifested as loops in the undirected structure of a BN. For example, when two parents, Y_1 and Y_2, of a node X_i have a common ancestor other than Z, or if one is an ancestor of the other, we can no longer claim that $P(Y_1|z)$ is independent of $P(Y_2|z)$ and can no longer use these quantities to fully account for $P(Y_1, Y_2|z)$. This problem is analogous to one in exact BN inference: belief propagation can be used to exactly and efficiently perform inference on polytrees (Rebane and Pearl, 1987), but not in general BNs due to dependencies that are introduced by loops.

Pearl (1988) introduced the method of cut-set conditioning to address this inference problem. Cut-set conditioning conditions on a set of variables (called the cut-set) that breaks these problematic loops. One can turn a BN into a polytree with a different parameter set for each assignment of the cut-set variables, perform exact inference on the polytrees, and then combine the results to answer the original inference query. One can use a similar approach with EDAPD: one would first find a cut-set that turns the graph structure into an ASCN under conditioning, and then perform the analysis for each cut-set assignment in a manner similar to that of EDAPD. The results of such an analysis would then be explanations of differences in the subsets defined by the cut-set assignments across the two data sets.

Implementation of cut-set conditioning and the application to the CEHC-PORT data immediately revealed that direct application of cut-set conditioning in this manner is problematic. The nature of the learned network is such that there are some well-connected regions, that yield fairly large (sometimes as many as 8) variables in a cut-set. Since the number of cut-set instantiations is exponential in the number of cut-set variables, this explodes the number of ASCN-explanations generated. The large amount of explanations

becomes difficult to perceive and understand. Another issue is that significant contributions to differences become less identifiable, as the effect of a contribution often becomes split across many smaller effects spread across the various cut-set instantiations.

There are other approaches to addressing the loop problem in BN inference, notably, the approach of converting the DAG into a different representation of the joint distribution, the factor tree. Inference is then performed on the factor tree by applying belief propagation in a straightforward manner (Lauritzen and Spiegelhalter, 1988). I discuss the possibility of taking a similar approach to explanation of differences in future work in Section 8.2.9.

Considering the problematic nature of cut-set based explanations, I decided to take a different approach to explanation, which focuses on maintaining intelligibility, even if at the cost of deviating from the tracing of inference. In this approach, as the explanation recurses from child to parent, it gains terms (value assignments of the child's other parents) on which the probabilities are conditioned throughout the traversal. For example, suppose variable A has parents B and C which themselves have parents in common. Then in the explanation of the marginal term $P(a)$ we would look at a joint probability term $P(a, b, c)$. In explaining the joint term, we would break up contributions to the ratio into contributions from the conditional term $P(a|b, c)$ and a joint among parents $P(b, c)$. The term $P(b, c)$ might, if C is not an ancestor of B, be broken up into $P(b) \times P(c|b)$, in which case the conditioning on b would be carried through the explanation regarding c. We could also, in principle, condition b on c instead, possibly yielding a different explanation. As described below, the approach remains sound regardless of the order in which we condition on variables, but there are advantages to conditioning variables that are generally lower in the topology of the graph on those that are higher.

6.4 DIFFERENCE EXPLANATION WITH CARRIED CONDITIONING

The DECC procedure follows the same general process as EDAPD, with the addition of a conditioning term that is carried along. To understand where this conditioning is coming from, and how to deal with it, look back to the decomposition that CPR yields: Starting at

a variable X_i, seeking to explain the difference in $P(x_{ik})$ across the two groups, the difference is broken down into contributions from each of the joint probability terms $P(x_{ik}, \pi_{ij})$ corresponding to the possible assignments of the parents of X_i to values x_{ik}. The difference in a joint probability term $P(x_{ik}, \pi_{ij})$ is in turn broken down to a contribution to the ratio from the conditional ratio term and the contributions of each parent ratio term $y_{tj}|y_{1j}, \ldots, y_{t-1,j}$. In CPR, we stopped the explanation at that stage. In EDAPD, we generated an explanation for each parent term by guaranteeing independence among the parents during model construction, which allowed us to recursively apply the procedure to each y_{tj}.

In DECC, the EDAPD procedure is modified to allow this recursive application without restricting the model structure. This is accomplished by accounting for additional conditioning from the very beginning. Suppose that our goal is to explain an observed difference in a probability of the form $P(x_{ik}|\mathbf{c})$ where $\mathbf{c}$ is the assignment of a subset $\mathbf{C} \subset \mathbf{X} \backslash \{X_i\}$ of variables to particular values. In a process analogous to that of CPR and EDAPD, we have that

$$P(x_{ik}|\mathbf{c}) = \sum_{j=1}^{J_i} P(x_{ik}, \pi_{ij}|\mathbf{c}) \tag{6.15}$$

which, letting $z_1 := \text{argmax}_z \, P(x_{ik}|\mathbf{c}, z)$, yields the following decomposition of the difference:

$$P(x_{ik}|\mathbf{c}, z_1) - P(x_{ik}|\mathbf{c}, \neg z_1) = \sum_{j=1}^{J_i} \left(P(x_{ik}, \pi_{ij}|\mathbf{c}, z_1) - P(x_{ik}, \pi_{ij}|\mathbf{c}, \neg z_1) \right) \tag{6.16}$$

where in the context of DECC we refer to the left-hand-side as the marginal difference term, and refer to the terms in the sum on the right-hand-side as the joint difference terms. Some parent assignments π_{ij} may be inconsistent with $\mathbf{c}$, meaning that $\mathbf{C}$ contains a parent of X_i and this parent's value in the assignment π_{ij} is different from its value in $\mathbf{c}$. Since this makes the probability $P(x_{ik}, \pi_{ij}|\mathbf{c}, Z)$ zero, these terms can be safely omitted as having no contribution. As in EDAPD and CPR, the remaining terms can be filtered by any of the tests in Section 4.4 and organized (grouped and sorted) for presentation purposes.

Let $z_2 := \text{argmax}_z \, P(x_{ik}, \pi_{ij}|\mathbf{c}, z)$. The ratio across groups for each joint term is broken up into the product of the conditional ratio term, and the product of parent ratio terms.

$$\frac{P(x_{ik}, \pi_{ij}|\mathbf{c}, z_2)}{P(x_{ik}, \pi_{ij}|\mathbf{c}, \neg z_2)} = \frac{P(x_{ik}|\pi_{ij}, \mathbf{c}, z_2)}{P(x_{ik}|\pi_{ij}, \mathbf{c}, \neg z_2)} \prod_{t=1}^{m} \frac{P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \mathbf{c}, z_2)}{P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \mathbf{c}, \neg z_2)} \tag{6.17}$$

There are two notable situations that warrant additional consideration regarding the parent terms. First, note that in the event that $Y_t \in \mathbf{C}$, since we previously eliminated inconsistencies, we have that $y_{tj} \in \mathbf{c}$, in which case $P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \mathbf{c}, z) = 1$ for any z, resulting in a ratio of 1 and no contribution from that term, meaning that it could be omitted from the explanation. For the remaining terms, depending on the network structure, the conditioning in $y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \mathbf{c}$ can often be simplified.

We can potentially find a smaller set of conditions $\mathbf{c}'$ which satisfies

$$P(y_{tj}|y_{1j}, \ldots, y_{t-1,j}, \mathbf{c}, z) = P(y_{tj}|\mathbf{c}', z) \qquad \forall z \ . \tag{6.18}$$

This simplified form would improve the clarity of presentation of the term to the user. Specifically, the set $\mathbf{c}'$ guarantees to satisfy (6.18) when it is the assignment of a set of variables $\mathbf{C}' \subset \mathbf{C}$ to the values they take in $\mathbf{c}$, such that $\mathbf{C}'$ d-separates Y_t from $\mathbf{C}$.

To find this subset of variables, I present the algorithm in Figure 14. Given a node Y and a set of nodes $\mathbf{W}$ in a BN, the algorithm finds $\mathbf{W}' \subset \mathbf{W}$ such that it d-separates Y from $\mathbf{W} \backslash \mathbf{W}'$ (the remaining nodes in $\mathbf{W}$). It is a breadth-first search of all trails that originate at Y and are d-connected by $\mathbf{W}$.

The search is performed as follows. In order to follow all possible trails, two queues (an 'up' queue and a 'down' queue) are kept. Initially, each queue contains only the node Y. The 'up' queue contains nodes from which search trails proceed upwards (towards parents), while the 'down' queue contains nodes from which search trails proceed downwards (towards children). Recall that a trail in a DAG is d-connected by $\mathbf{W}$ iff for every triple of consecutive nodes A, B, C in the trail, whenever B is a collider (the arc directions are $A \rightarrow B \leftarrow C$) it holds that $B \in \mathbf{W}$ (Geiger et al., 1990b). When we pop a node U from the 'up' queue, we consider each parent P of U. If the parent is in $\mathbf{W}$, the parent is added to the result set. Additionally, if $P \in \mathbf{W}$, the trail terminates at the parent (since P d-separates both $U \leftarrow P \leftarrow \cdot$ trails and $U \leftarrow P \rightarrow \cdot$ trails). If $P \notin \mathbf{W}$, neither $U \leftarrow P \leftarrow \cdot$ trails nor $U \leftarrow P \rightarrow \cdot$ trails are d-separated, and hence P is added to both the 'up' queue and the 'down' queue. When we pop a node D from the 'down' queue, we consider each child C of D. If the child is in $\mathbf{W}$, the child is added to the result set. Additionally, if the child is in $\mathbf{W}$, then it d-separates $D \rightarrow C \rightarrow \cdot$ trails, and d-connects $D \rightarrow C \leftarrow \cdot$ trails. We therefore add

```
1: function SIMPLIFY(Y,W)
2:     Q_u ← (Y)                                  ▷ The queue of 'up' nodes.
3:     Q_d ← (Y)                                  ▷ The queue of 'down' nodes.
4:     V_u ← ∅                                    ▷ The set of visited 'up' nodes.
5:     V_d ← ∅                                    ▷ The set of visited 'down' nodes.
6:     W' ← ∅                                     ▷ The set that will contain the result.
7:     while Q_u ≠ ∅ ∨ Q_t ≠ ∅ ∧ |W'| < |W| do
8:         if Q_u ≠ ∅ then
9:             U ← POP(Q_u)
10:            if U ∉ V_u then
11:                V_u ← V_u ∪ {U}
12:                P ← PARENTS(U)
13:                W' ← W' ∪ (W ∩ P)
14:                PUSH(P\W, Q_u)                 ▷ Follow U ← P ← · trails.
15:                PUSH(P\W, Q_d)                 ▷ Follow U ← P → · trails.
16:        if Q_d ≠ ∅ then
17:            D ← POP(Q_d)
18:            if D ∉ V_d then
19:                V_d ← V_d ∪ {D}
20:                C ← CHILDREN(D)
21:                W' ← W' ∪ (W ∩ C)
22:                PUSH(C ∩ W, Q_u)               ▷ Follow D → C ← · trails.
23:                PUSH(C\W, Q_d)                 ▷ Follow D → C → · trails.
24:    return W'
```

Figure 14: Algorithm for simplifying a set of conditioned variables using d-separation.

C only to the 'up' queue. Conversely, if the child is not in $\mathbf{W}$, then it d-connects $D \rightarrow C \rightarrow \cdot$ trails, and d-separates $D \rightarrow C \leftarrow \cdot$ trails, and we add it only to the 'down' queue.

To avoid unnecessarily re-visiting nodes, a set of nodes visited for each queue is maintained and checked every time a node is popped from the corresponding queue. The search terminates either when all nodes that can be reached by d-connected trails are visited, or when all members of $\mathbf{W}$ enter the result set $\mathbf{W}'$. When the search terminates, $\mathbf{W}'$ contains all members of $\mathbf{W}$ that are d-connected to Y by $\mathbf{W}$; therefore, Y is d-separated by $\mathbf{W}'$ from $\mathbf{W} \backslash \mathbf{W}'$.

This algorithm gives us $\mathbf{c}'$ that satisfies (6.18). We can then apply a significance test from Section 4.4 to the event $y_{tj}|\mathbf{c}'$, and, if the term is found to be significant, recursively apply DECC to $y_{tj}|\mathbf{c}'$.

The SIMPLIFY function can also be used to reduce the number of conditions in the $x_{ik}|\pi_{ij}, \mathbf{c}$ term by applying it to find a subset of conditions $\mathbf{a}$ s.t. $\pi_{ij} \subset \mathbf{a} \subset \pi_{ij} \cup \mathbf{c}$ and

$$P(x_{ik}|\pi_{ij}, \mathbf{c}, z) = P(x_{ik}|\mathbf{a}, z) \qquad \forall z \ . \tag{6.19}$$

Figure 15 summarizes the process. Since the aim of the procedure is to use it to explain a difference in the marginal probability of a variable assignment x_{ik}, DECC is initially called with $\mathbf{c} = \emptyset$. From that point onward, the conditioning term $\mathbf{c}$ in each step is determined in the recursive call to DECC. While it is theoretically possible for DECC to be called a number of times exponential in the number of variables, pruning by the statistical significance tests make such an event unlikely in practice.

Note that both EDAPD and CPR are special cases of DECC: DECC gives the CPR explanation if we set the CLINICALSIGNIFICANCETEST test for the parent marginal to always fail, and it gives the EDAPD explanation when the ancestry graph of X_i is an ASCN.

6.4.1 Clinical data case study

This section presents a case study of applying the DECC explanation method to the CEHC-PORT clinical data described in Section 2.4. Using the multi-model approach, a BN was learned with order constraints on the variables as described in Section 6.1.1.

1: **procedure** DECC(x_{ik},$\mathbf{c}$)
2: $\quad$ REPORT($x_{ik}|\mathbf{c}, Z$)
3: $\quad$ Let $z_1 := \text{argmax}_z\, P(x_{ik}|\mathbf{c}, z)$
4: $\quad$ **for** $j \in \{1, \ldots, J_i\}$ **do**
5: $\quad\quad$ **if** CONSISTENT($\pi_{ij}, \mathbf{c}$) **then**
6: $\quad\quad\quad$ **if** CLINICALSIGNIFICANCETEST($x_{ik}, \pi_{ij}|\mathbf{c}$) **then**
7: $\quad\quad\quad\quad$ REPORT($x_{ik}, \pi_{ij}|\mathbf{c}, Z$)
8: $\quad\quad\quad\quad$ $\mathbf{a} \leftarrow \pi_{ij} \cup \mathbf{c}$
9: $\quad\quad\quad\quad$ $\mathbf{A} \leftarrow$ SIMPLIFY($X_i, \mathbf{A}$)
10: $\quad\quad\quad\quad$ REPORT($x_{ik}|\mathbf{a}, Z$)
11: $\quad\quad\quad\quad$ **for** $t \in \{1, \ldots, m\}$ **do**
12: $\quad\quad\quad\quad\quad$ **if** $Y_t \notin \mathbf{C}$ **then**
13: $\quad\quad\quad\quad\quad\quad$ $\mathbf{c}' \leftarrow \mathbf{c} \cup \{y_{1j}, \ldots, y_{t-1,j}\}$
14: $\quad\quad\quad\quad\quad\quad$ $\mathbf{C}' \leftarrow$ SIMPLIFY($Y_t, \mathbf{C}'$)
15: $\quad\quad\quad\quad\quad\quad$ REPORT($y_{tj}|\mathbf{c}', Z$)
16: $\quad\quad\quad\quad\quad\quad$ **if** CLINICALSIGNIFICANCETEST($y_{tj}|\mathbf{c}'$) **then**
17: $\quad\quad\quad\quad\quad\quad\quad$ DECC(y_{tj},$\mathbf{c}'$)

Figure 15: The recursive DECC analysis procedure.

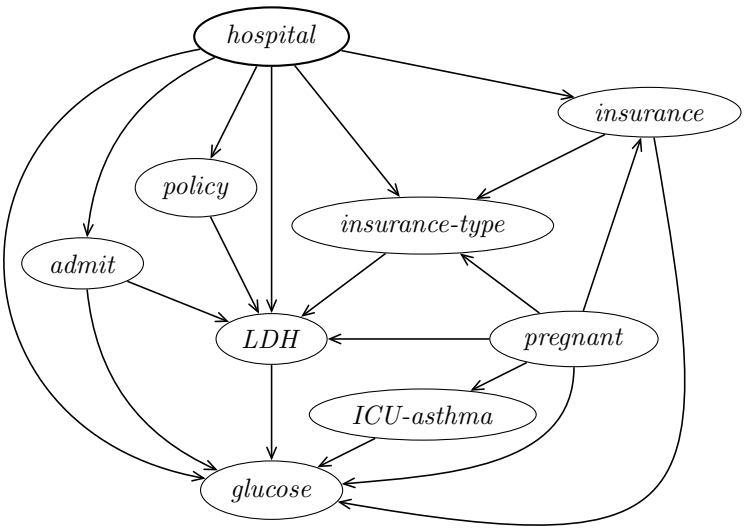

Figure 16: The fragment of the multi-model BN learned from the CHEC-PORT data which is relevant to the explanation of the marginal probability difference in the *glucose* variable.

Clinical significance was determined by using the tests from Section 4.4, specifically, the absolute difference test for joint difference terms at a threshold of $\delta = 0.1$ and the absolute ratio test for parent ratio terms at a threshold of $\varepsilon = \log 1.1$. These thresholds were picked because they were expected to provide an informative but manageable level of detail for the purposes of this case study.

Let us examine the *glucose* variable. The model indicates a significant difference in the marginal probability that a patient has a glucose level of *glucose* = 70–110, a probability of 0.488 in hospital A and 0.593 in hospital B, yielding a difference of 0.105. Figure 16 shows the part of the multi-model BN that is relevant to explaining this difference.

The *glucose* variable has six parents in the network, including the *hospital* node. The other parents are *pregnant*, *insurance*, *admit* (admission source), *LDH* (lactic dehydrogenase), and *ICU-asthma* (ICU admission due to asthma in the past year). The largest significant joint difference term that contributes towards the marginal difference corresponds to the parent assignment [*admit* = missing, *pregrant* = no, *ICU-asthma* = no, *insurance* = yes, *LDH* $<$ 170]. Joint probability of *glucose* = 70–110 and this parent assignment is

0.268 in hospital A and 0.382 in hospital B, yielding a difference of 0.114. The ratio of these probabilities is 1.427. The conditional term has no contribution to the ratio, but the association between the glucose level and the parent assignment is strong. The conditional probability of *glucose* = 70–110 given this parent assignment is 0.929.

Note that the conditional independence between *glucose* and *hospital* here is an instance of context-specific independence due to the multi-model network construction. There is an arc from *hospital* to *glucose* in the network structure, and, for example, for a different parent assignment, where *admit* = ER, we do see a difference in the conditional probabilities (0.288 in A and 0.347 in B).

The contributions of the individual parent terms to the joint ratio term above are: No contribution for *pregnant* = no, 1.001 (not clinically significant) for *ICU-asthma* = no given *pregnant* = no, 1.087 (not clinically significant) for *insurance* = yes given *pregnant* = no (conditioning simplification eliminated the conditioning on *ICU-asthma*), 1.09 (not clinically significant) for *admit* = missing, and 1.156 (clinically significant) for *LDH* = $<$ 170 given *pregnant* = *no*, *admit* = missing, and *insurance* = yes. The probabilities for the *LDH* term are 0.838 in hospital A and 0.968 in hospital B, yielding a difference of 0.13.

LDH has five parents: *insurance-type*, *policy*, *pregnant*, *admit*, and *insurance*. We continue to examine the joint terms that contribute to this difference. Note that three of the parents of *LDH* have their states determined by the conditioning that was carried along. The joint term that accounts for most of the difference has the parent assignment *insurance-type* = Q[2], *policy* = no given *pregnant* = *no*, *admit* = missing, and *insurance* = yes. The joint probability of *LDH* $<$ 170 and this parent assignment is 0.001 in hospital A and 0.953 in hospital B. The contribution from this joint term to difference in the *LDH* term is 0.952, and it is balanced out by a significant term that contributes 0.754 against the difference (with *insurance-type* = Q').

The joint ratio term for [*LDH* $<$ 170, *insurance-type* = Q, *policy* = no given *pregnant* = *no*, *admit* = missing, and *insurance* = yes] is 717.412. The conditional probability of *LDH* $<$ 170 given the parent assignment is 0.5 in hospital A and 0.976 in hospital B. The contribution of the conditional term to the ratio is 1.951.

[2]I encode the actual insurance type to anonymize the data

Due to the carried conditioning, only two of the five parent terms can, in principle, contribute to the joint ratio term. The contributions from those parent terms are broken down as follows: 1.001 (not significant) from *policy* = missing, and 367.304 (accounts for most of the ratio) from *insurance-type* = Q given *pregnant* = no and *insurance* = yes. The probabilities for *insurance-type* = Q given *pregnant* = no and *insurance* = yes are 0.003 in hospital A and 0.978 in hospital B. The only parents of *insurance-type* are *hospital*, *pregnant*, and *insurance*, meaning that this contribution is an elementary difference in the model.

There are two strong relationships that play a central part in the explanation of differences. One is the strong link between LDH and glucose. We see a strong association between normal glucose levels (70–110) and normal to low LDH levels (< 170). Examining the data further, for other values of *glucose* reveals that similarly, high glucose levels correlate with high LDH levels. This relationship makes sense from a medical standpoint since LDH and glucose participate in related metabolic pathways.

The second strong relationship that contributes to the difference is not biological in nature at all, but has to do with a difference between the economic environments surrounding the hospitals: most patients in hospital A have insurance type Q, while most patients in hospital B have a different insurance type. I conjecture that the reason for the appearance of this variable in the explanation, is that this strong relationship has made the *insurance-type* variable act as a proxy for the *hospital* variable in the $\mathcal{M}_{\cup}$ model in the multi-model. This inclusion then propagated to the synthesized model. This observation suggests that multi-model approach can be sensitive to this sort of effect from variables that highly correlate with the group division.

Variables that are related to information about a patient's health insurance can also be associated with an individual patient's medical condition. The presence or absence of insurance and type of insurance can act as a proxy for indicating the patient's socio-economic status, which would be strongly related to their general ability and tendency to get treatment and their level of access to preventive healthcare.

7.0 BUILDING AN INTERACTIVE GRAPHICAL USER INTERFACE

As we see, the explanation generated by the algorithms above may be fairly complex objects, even for medium-sized data (dozens to hundreds of variables). This makes static presentation of the explanation somewhat cumbersome. This chapter discusses the development of an interface to facilitate interactive exploration of these explanations, presents a prototype that I have implemented, and discusses some next steps that would be helpful to turn the system into an application for general use.

7.1 DESIGN CONSIDERATIONS

Let us first consider the most basic manner in which an explanation can be presented, the static textual report, since it is the limitations of this method that motivate the need for an interactive system for presenting the explanation.

Consider the case study from Section 6.2.3. The following is an excerpt of the report that EDAPD generated the *CRI* variable, using the Bayesian difference test (4.30) with $\delta = 0.01$ and $\alpha = 0.05$ for filtering joint terms:

```
*** Analyzing CRI ***
*** Difference threshold 0.01 ***
*** Significance level 0.05 ***

   (Q)P( CRI=yes | HOSPITAL ):
   0.081(A) / 0.052(B) = 1.547
   C.I: (1.324,1.796)

   (D)P( CRI=yes | HOSPITAL ):
   0.081(A) - 0.052(B) = 0.028
```

```
C.I: (0.017,0.041)

   === Terms contributing towards the difference ===

      (D)P( CRI=yes, HTNA: yes | HOSPITAL ):
      0.065(A) - 0.032(B) = 0.033
      C.I: (0.021,0.047)

      (Q)P( CRI=yes, HTNA: yes | HOSPITAL ):
      0.065(A) / 0.032(B) = 2.043
      C.I: (1.65,2.489)

         Conditional part:
         P( CRI=yes | HTNA: yes | HOSPITAL ):
         0.172(A) / 0.172(B) = 1
         C.I: (1,1)

      === Terms contributing towards the quotient ===

         (Q)P( HTNA=yes | HOSPITAL ):
         0.378(A) / 0.185(B) = 2.043
         C.I: (1.65,2.489)

         (D)P( HTNA=yes | HOSPITAL ):
         0.378(A) - 0.185(B) = 0.193
         C.I: (0.132,0.253)

            === Terms contributing towards the difference ===

               (D)P( HTNA=yes, AGEPRES6: 30-44 | HOSPITAL ):
               0.129(A) - 0.089(B) = 0.041
               C.I: (0.011,0.069)

               (Q)P( HTNA=yes, AGEPRES6: 30-44 | HOSPITAL ):
               0.129(A) / 0.089(B) = 1.456
               C.I: (1.115,1.884)

                  Conditional part:
                  QP( HTNA=yes | HOSPITAL ):
                  0.487(A) / 0.439(B) = 1.11
                  C.I: (0.906,1.354)

               === Terms contributing towards the quotient ===

                  (Q)P( AGEPRES6=30-44 | HOSPITAL ):
                  0.266(A) / 0.202(B) = 1.312
                  C.I: (1.102,1.561)
```

```
   (D)P( AGEPRES6=30-44 | HOSPITAL ):
   0.266(A) - 0.202(B) = 0.063
   C.I: (0.022,0.106)

      HOSPITAL is the only parent of AGEPRES6

=== Terms contributing against the quotient ===

(D)P( HTNA=yes, AGEPRES6: 45-59 | HOSPITAL ):
0.054(A) - 0.015(B) = 0.039
C.I: (0.016,0.065)

(Q)P( HTNA=yes, AGEPRES6: 45-59 | HOSPITAL ):
0.054(A) / 0.015(B) = 3.61
C.I: (1.861,6.832)

   Conditional part:
   QP( HTNA=yes | HOSPITAL ):
   0.467(A) / 0.107(B) = 4.356
   C.I: (2.353,8.139)

=== Terms contributing towards the quotient ===

=== Terms contributing against the quotient ===

(D)P( HTNA=yes, AGEPRES6: 60-74 | HOSPITAL ):
0.081(A) - 0.015(B) = 0.065
C.I: (0.022,0.109)

(Q)P( HTNA=yes, AGEPRES6: 60-74 | HOSPITAL ):
0.081(A) / 0.015(B) = 5.342
C.I: (2.151,12.99)

   Conditional part:
   QP( HTNA=yes | HOSPITAL ):
   0.381(A) / 0.051(B) = 7.505
   C.I: (3.061,17.606)

=== Terms contributing towards the quotient ===

=== Terms contributing against the quotient ===

   (Q)P( AGEPRES6=60-74 | HOSPITAL ):
   0.297(B) / 0.211(A) = 1.405
   C.I: (1.181,1.674)
```

```
                  (D)P( AGEPRES6=60-74 | HOSPITAL ):
                  0.297(B) - 0.211(A) = 0.086
                  C.I: (0.043,0.128)

                     HOSPITAL is the only parent of AGEPRES6

               (D)P( HTNA=yes, AGEPRES6: 18-29 | HOSPITAL ):
               0.033(A) - 0.002(B) = 0.031
               C.I: (0.018,0.046)

               (Q)P( HTNA=yes, AGEPRES6: 18-29 | HOSPITAL ):
               0.033(A) / 0.002(B) = 13.799
               C.I: (4.32,129.642)

                  Conditional part:
                  QP( HTNA=yes | HOSPITAL ):
                  0.214(A) / 0.027(B) = 8.013
                  C.I: (2.496,71.793)

               === Terms contributing towards the quotient ===

                  (Q)P( AGEPRES6=18-29 | HOSPITAL ):
                  0.154(A) / 0.09(B) = 1.722
                  C.I: (1.346,2.349)

                  (D)P( AGEPRES6=18-29 | HOSPITAL ):
                  0.154(A) - 0.09(B) = 0.065
                  C.I: (0.035,0.102)

                     HOSPITAL is the only parent of AGEPRES6

               === Terms contributing against the quotient ===

         === Terms contributing against the difference ===

   === Terms contributing against the quotient ===

=== Terms contributing against the difference ===
```

The main issue with such a report is that even with the filtering of less significant terms, and even in this relatively simple case, it is long. Yet, all the details are necessary for understanding the explanation. In my presentation of the case study, I verbally summarized this report into prose. Even such a summary is problematic, not only because it hides details in an attempt at brevity and clarity, but also because on its own it reads almost like a riddle

that a reader needs to parse to work out the relationships between the variables. With the aid of the relevant network structure (Figure 13), the nature of the relationships between variables is much clearer, but the graph itself does not provide essential information regarding the degree to which the distributional differences across hospitals in each age group and in hypertension contribute to the observed difference in rates of CRI.

7.1.1 Importance and challenges of the graph visual

These considerations suggest that the a good explanation presentation should include the graphical structure of the BN, and supplement it with the details that are not represented in the graphical structure. This turns out to be a nontrivial challenge when considering that networks such as the one learned from the PORT Data are difficult to lay out in any intelligible way because of how highly connected they are. Moreover, adding elements to the visualization of such a network would exacerbate the issue. Interactivity presents an elegant solution to the problem of viewing a complicated BN: we can limit the portion of the BN explained to those nodes that are relevant to the explanation, in much the same way that Figure 13 is a relevant fragment from a network of over a hundred nodes, specifically, it is the ancestry graph of the CRI node. While that is not the case in the EDAPD case study, the ancestry graph of a node can still be quite large and complex. I will return to this issue.

It is very tempting to seek an approach that does not augment a textual explanation with a graph structure, but rather represents all the information in the context of the graph structure.

The main issue with incorporating the information within an explanation into the graph structure is that the probabilities considered in the explanation are dependent on particular values of variables. It is not entirely accurate to say, for example, that "hypertension differs across the hospitals when controlling for age" because we see a significant difference when conditioning on some age groups, but not on others. This level of detail cannot be captured in an intuitive way by techniques such as coloring or styling the edges and nodes in the BN graph, since there are many different, possibly even opposing, effects that an edge or node would be associated with (e.g., hypertension being more prevalent in hospital A for one age

group, in hospital B in another, and having equal prevalence across hospitals in another).

These issues lead to a presentation design where the BN structure is present, but it is augmented with an independent display of the explanation which is suited for the level of detail that we wish to convey.

7.1.2 Explanations are tree-structured

While text can indeed provide the level of detail desired, its sequential nature makes it less than optimal for our task. Conceptually, the explanation is not a flat sequence of statements, but is tree-structured. Figure 17 shows an abstraction of this structure. The display above uses indentation to hint at the structure. However, for longer explanations, it may become difficult to keep track of the structure with indentation alone, without a more explicit representation of the structure. Let us take the case study above and augment it with the explicit tree structure.

Figure 18 shows a fragment of this explanation put into its explicit tree structure. We can see that the tree visualization helps see how each block, which corresponds to a decomposition term, relates to its parent block, which is the term to which it contributes. The explanation length, however, is still an issue. Even though Figure 18 is only a fragment of the full explanation generated, it still takes considerable effort to parse and understand it in its entirety. This is another aspect of the explanation presentation that can benefit from interactivity. With an interactive tree that allows for expanding and collapsing branches, the user can get a better sense of the overall picture.

In summary, effective explanation presentation challenges us to handle the size of the explanation as well as stay well-oriented within it, with awareness of the context of each explanation atom. These challenges have led me to adopt a design that has two main components: an interactive tree-structured display of the explanation, and a display of the BN graph structure that is relevant to the explanation atom that a user is examining at a given moment.

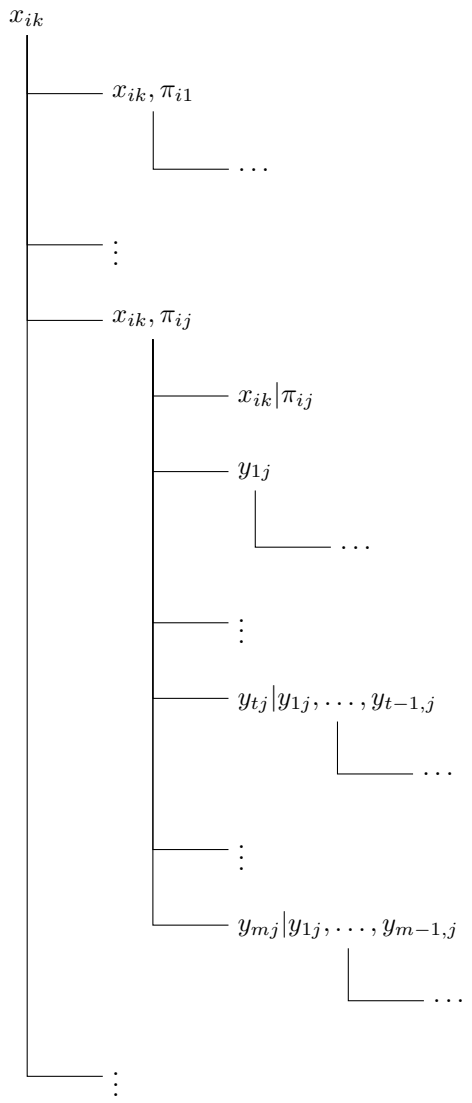

Figure 17: Abstract illustration of the tree structure of an explanation.

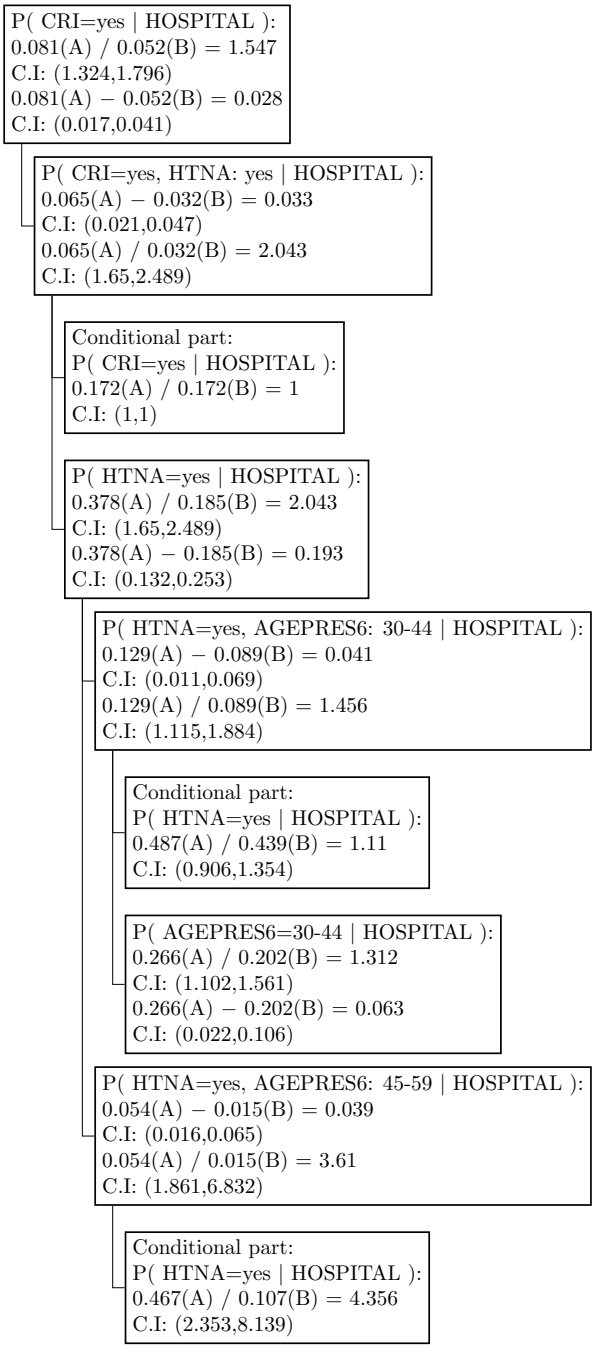

Figure 18: A tree representation of a part of the EDAPD explanation for the CRI case study.

7.2 PROTOTYPE

The design consideration above guided the creation of a prototype application for interactive explanation presentation. I implemented the prototype in the Java programming language, using jSMILE[1] for BN inference and the JUNG library[2] for graph visualization. In this section I discuss the appearance and function of the prototype.

Figure 19 shows the initial view that a user sees when loading data to explain. There are two panes, the right-hand side is a tree-structured view of the explanation, where the top level tree nodes correspond to the variables that appear in the data. The left-hand side is used to display relevant graph structure, it is initially empty. As I show below, the selection of explanation elements on the right-hand pane triggers the display of graph elements on the left-hand side.

One important feature to note is that the variables are sorted in the list by descending magnitude of the differences of marginal probability across groups (in this case, across hospitals). That is, by descending order of

$$\max_{k=1}^{K_i} |P(x_{ik}|Z=2) - P(x_{ik}|Z=1)| \ . \tag{7.1}$$

This implies that the first variable listed will always be Z, and the rest of the variables follow, each having a larger marginal difference than the next, with variables that have no difference across the groups at the bottom. For completeness, I show in Figure 20 what one can see by expanding the HOSPITAL (Z) node, which is merely tautological ($P(z|z) = 1$ and $P(z|\neg z) = 0$).

The left-hand side of Figure 20 demonstrates the triggering of the graph display. Selecting the CLDH variable in the tree on the right-hand side triggered the display of all the parents of CLDH in the BN, as well as any direct arcs among the parents. Note that the absence of arcs in this display does not preclude the existence of indirect dependence among the parents mediated by variables that are not presently displayed.

[1] jSMILE is the Java API for the SMILE reasoning engine for graphical probabilistic models contributed to the community by the Decision Systems Laboratory, University of Pittsburgh and available at http://genie.sis.pitt.edu/.

[2] JUNG—the Java Universal Network/Graph Framework—is a software library for the modeling, analysis, and visualization of graphs, available at http://jung.sourceforge.net/.

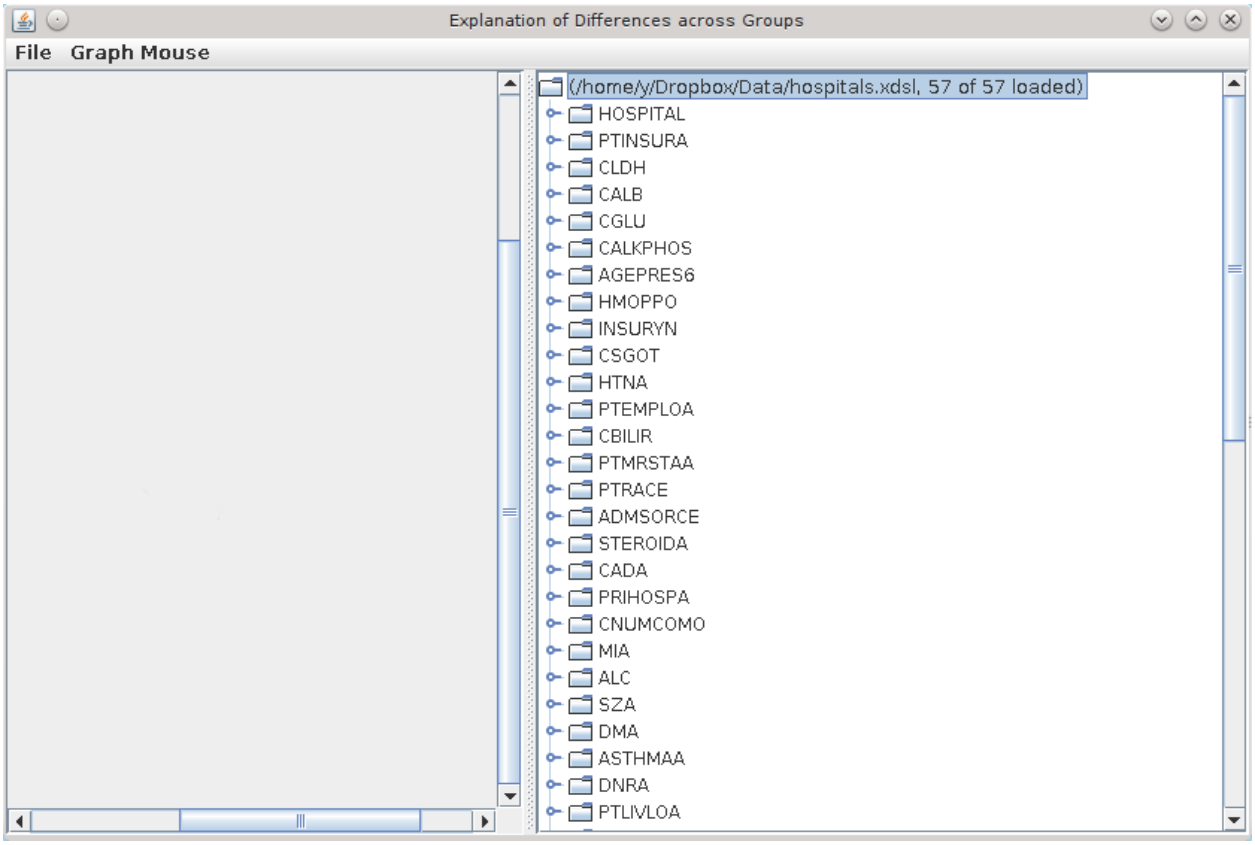

Figure 19: The prototype, initial view.

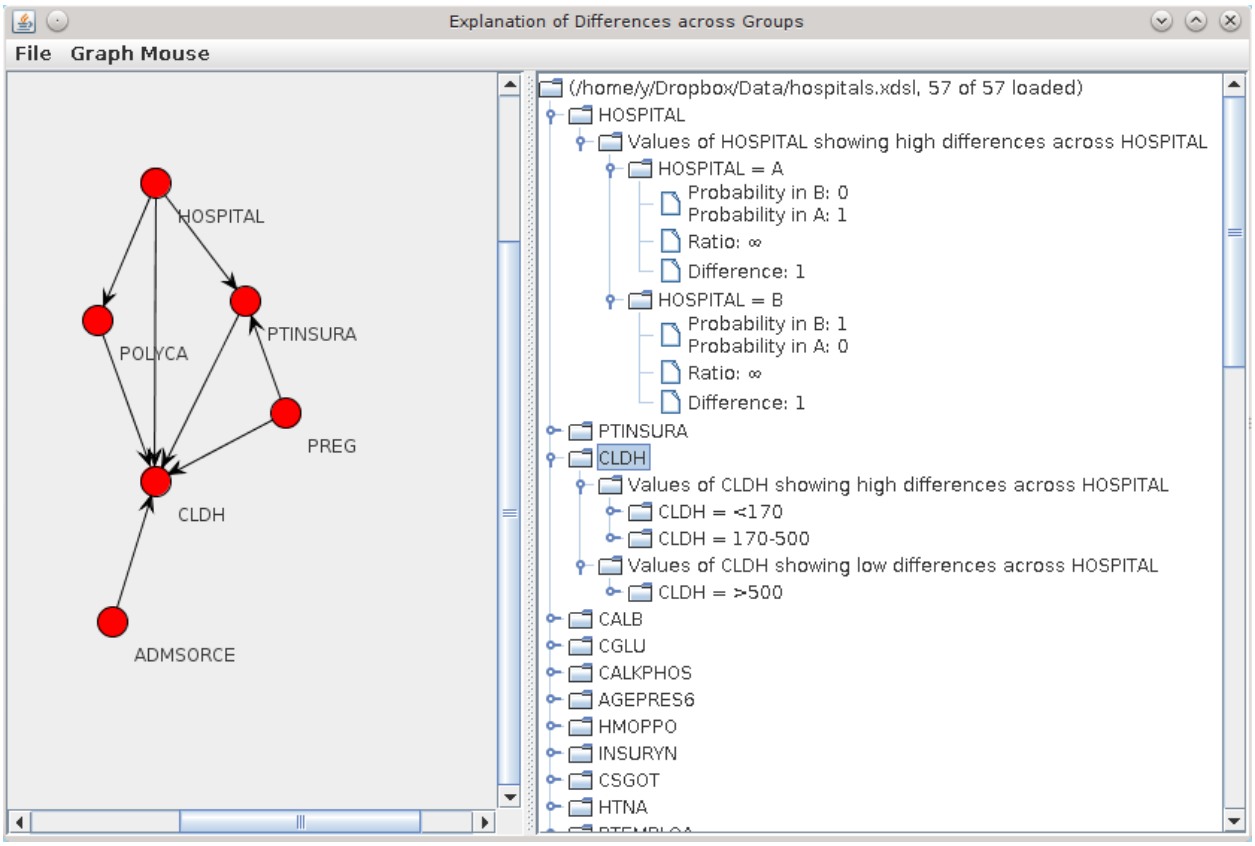

Figure 20: The prototype with CLDH selected and expanded to list its values, and with HOSPITAL fully expanded.

Figure 20 also shows that expanding a variable in the tree, in this case, the CLDH node, reveals two child nodes that group the values that the node (CLDH) can take into values showing 'high' and 'low' differences. I use a clinical significance test to determine whether a difference is 'high' or 'low'. In the prototype, x_{ik} is considered as showing a 'high' difference if

$$|P(x_{ik}|Z = 2) - P(x_{ik}|Z = 1)| > \delta \tag{7.2}$$

with $\delta = 0.1$, and considered as showing a 'low' difference otherwise. By expanding these grouping nodes in the explanation tree we see that CLDH takes three values, < 170, 170-500, and > 500. Note that within the groupings, the values are also sorted in descending magnitude of the absolute difference of marginal probabilities across groups.

Selecting a particular value of a variable, e.g. selecting '< 170' for CLDH, updates the graph labels on the left-hand pane to indicate the selection, as shown in Figure 21. The figure also shows that expanding the node reveals details about the explanation term for 'CLDH < 170,' specifically, probabilities in each group, and the ratio and difference between the probabilities. Since the explanation procedure is based on the decomposition of $P(x_{ik}|Z)$ into a sum of $P(x_{ik}, \pi_{ij}|Z)$ terms, I made the difference term expandable. Expanding the difference node reveals two nodes that represent groupings of those terms that make up the difference. There is one group that lists terms accounting for most of the difference, and a second group for all remaining terms. Terms are included in the former group as follows: the x_{ik}, π_{ij} terms are added one by one to the group of "terms accounting for most of the difference" in descending order of magnitude until the sum of the included terms accounts for more than half of the difference in $P(x_{ik}|Z)$. I found that in practice, the difference is often dominated by only a few terms. A grouping such as the one in Figure 22 seems to be common. This makes it very easy for a user to concentrate only on the terms that drive the difference of interest.

Note the left-hand pane in Figure 22 that selecting particular x_{ik}, π_{ij} term updates the graph labels to show the particular assignment of parents to values that is associated with the selection. Expanding a x_{ik}, π_{ij} node in the explanation tree node shows the probabilities associated with it in each group and the difference and ratio of probabilities. The next level of explanation is to break down the ratio associated with the x_{ik}, π_{ij} term into the contributions

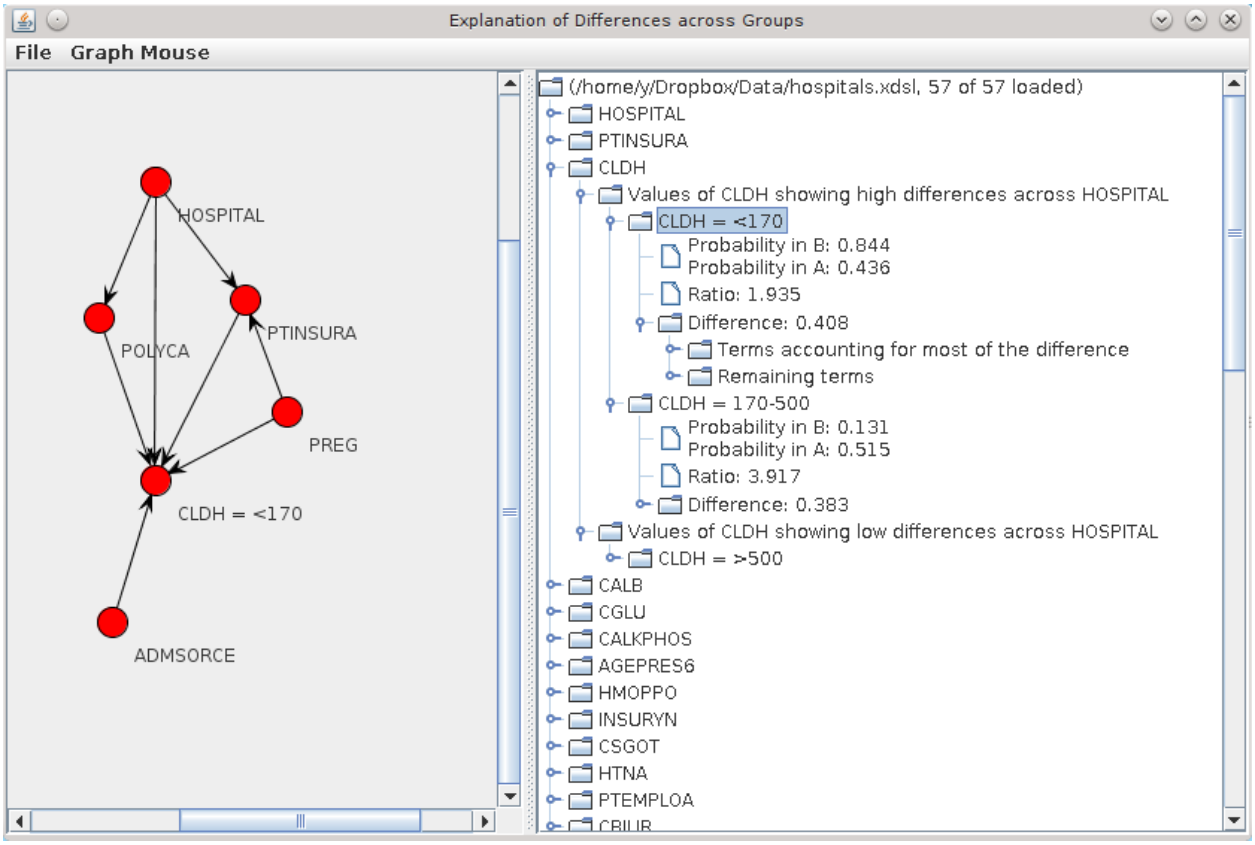

Figure 21: The prototype with CLDH< 170 selected and expanded.

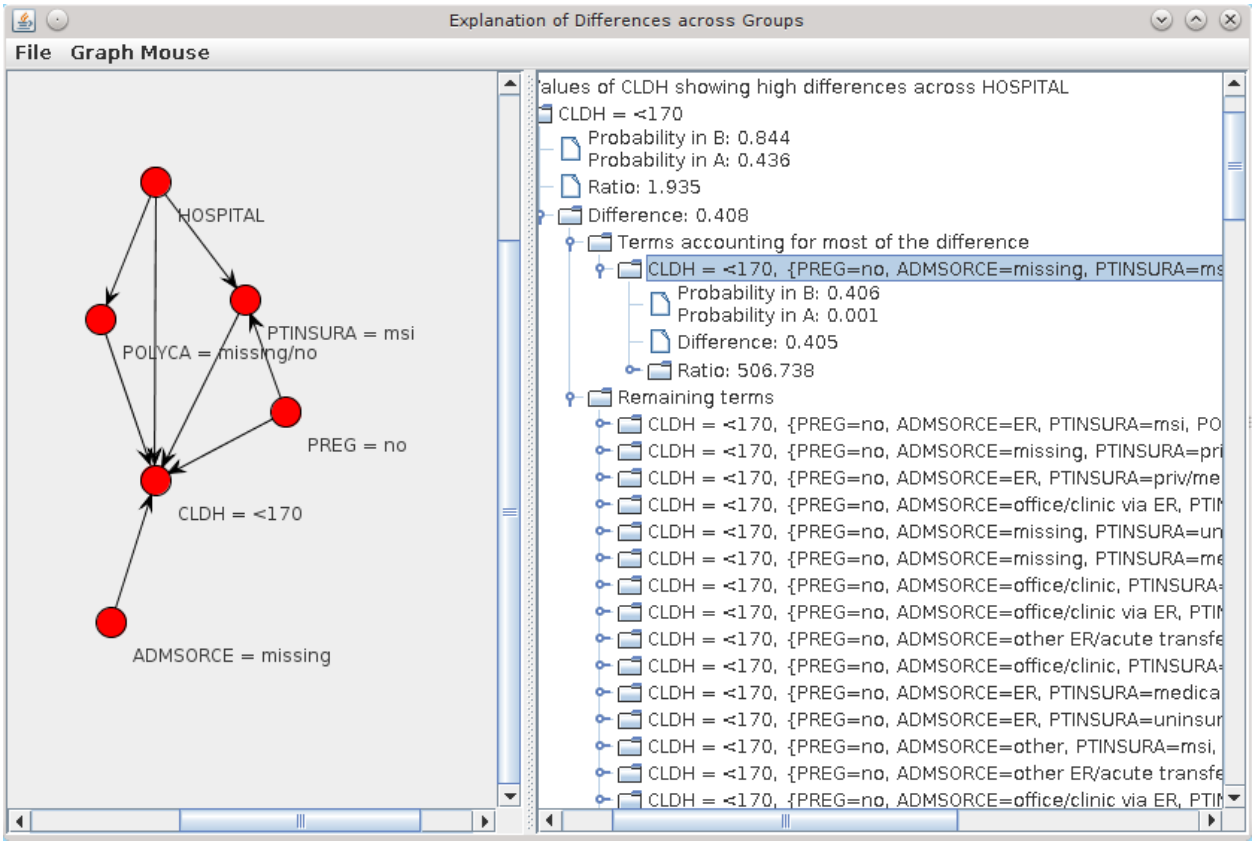

Figure 22: The prototype with a difference node expanded and a x_{ik}, π_{ij} node selected.

to the ratio from a conditional term $x_{ik}|\pi_{ij}$ and a joint term over the parent assignment π_{ij}. I show in Figure 23 that the probabilities, difference, and ratio of probabilities is reported for the conditional term. The contribution of all the parents jointly to the ratio is reported, and expanding the π_{ij} node reveals individual parent terms. Each parent term is analogous to the x_{ik} terms that we encountered when we selected a specific value after expanding the variable nodes that are at the top of the explanation tree. This stage is analogous to the recursion step in DECC, and just as we do in DECC, we list the parents in topological order and, when necessary, add a conditioning term to account for dependence among parents.

When these individual parent terms are selected, the graph structure is updated to include their immediate parents, as in Figure 24, which shows the graph update in response to the selection in Figure 25. Figure 25 shows that expanding the recursive nature of the explanation: expanding the individual parent term reveals an expandable difference term, which reveals terms contributing to the difference, each of which has its own ratio decomposition, and so on.

7.3 DISCUSSION

In this section I discuss the strengths and weaknesses of the prototype. I collected preliminary feedback from a clinical research expert regarding the prototype. I incorporate this feedback along with more general observations, as well as notable details regarding implementation. I discuss ideas regarding what future development would be required to evolve the prototype into an application that can become a part of a clinical researcher's standard analysis toolbox.

A technical strength demonstrated in the prototype is that by interactively revealing only those components of an explanation that a user chooses to explore, the computational burden of generating an explanation is reduced. While the generation algorithms in Chapter 6 can potentially require exploring the explanation tree fully, creating the necessity for pruning to reduce this burden, an interactive display of the explanation tree can limit the generation only to those parts of the tree that are visible on-demand. Indeed, in the prototype implementation, the expansion of a node in the explanation tree triggers the generation

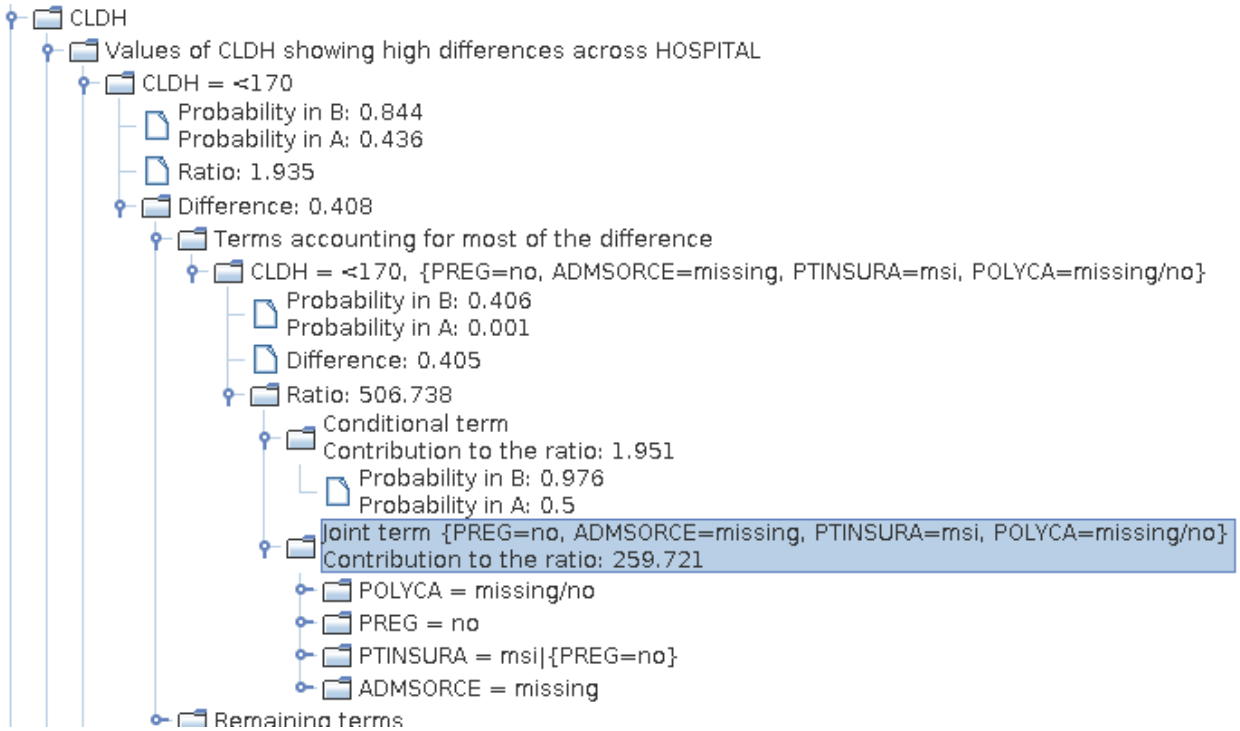

Figure 23: The prototype (explanation tree only) with a π_{ij} term highlighted and expanded.

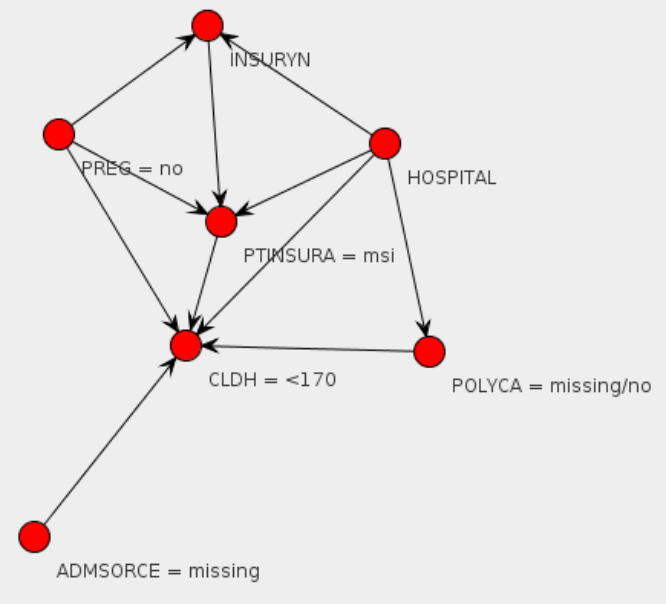

Figure 24: The prototype (graph only) showing the addition of the INSURYN node as a result of examining the PTINSURA node.

of additional explanation layers under that node, thus deferring, and potentially reducing, the amount of computation that needs to take place.

Feedback about the prototype interface was overall positive. As hypothesized in Section 7.1, the presence of a graph visualization was helpful and enhanced the clarity of presentation of the relationships found in the data. However, while it was apparent that the graph visualization responded to navigation through the explanation tree, the exact nature of the updates was not immediately obvious. For future development, therefore, it would help to add elements that would further clarify the correspondence between e.g. the variable the differences of which are being explained and its place in the graph, or the value assignments corresponding to a given explanation term and the appearance of those assignments in the graph visualization. This sort of correspondence can be pointed out by displaying an annotated tutorial example to the user as an aid to using the software. Another, more streamlined approach would be to highlight the correspondence using visual cues, such as using matching colors to match terms in the explanation tree to terms in the graph visu-

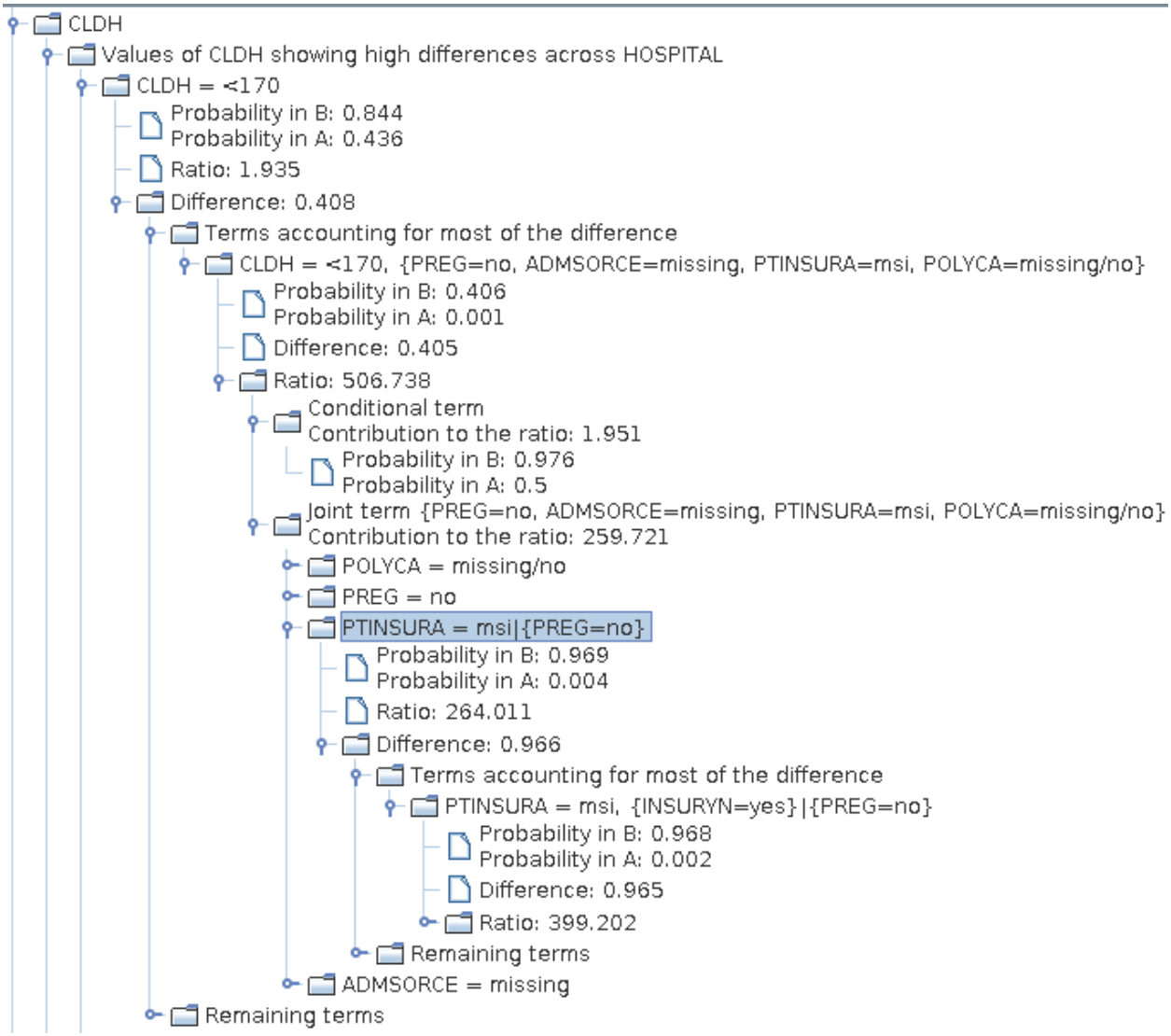

Figure 25: The prototype (explanation tree only) with an individual parent term selected and expanded.

alization. One could also mark changes in the graph (the appearance or disappearance of nodes, the changes of labels) with visual cues. When a user sees a response in a graph to a navigation action in the explanation tree, such marks that draw attention to the change in the graph will help associate the change in the graph with a step in the explanation tree.

Direct feedback revealed that some other features present in the prototype are perhaps not clear to a user first encountering it. One such feature appears in Figure 19, the very first screen. A question that came up is whether there is an easy way to see all the variables that show differences at once. The answer is that it is in fact very easy, since the variables are sorted in order of the magnitudes of the differences, and the list of all variables that have different distributions is indeed at the top of the variable list the user first sees.

This indicates that the sorting needs to be made more apparent. One option is to include instructions explaining the sorting for the user. A more intuitive approach is to show the quantity by which the variables are sorted. Another feature that may improve utility in this area is that of marking all variables that fall above or below a certain threshold. Yet another related feature, which was suggested by the expert, is that of including an option to change the measure by which the variables are sorted, switching between magnitude of differences and magnitude of ratios.

This applies not only to sorting the top-level list of variables, but also the sorting and grouping of the additive x_{ik}, π_{ij} terms in the difference decomposition and the multiplicative terms in the ratio decomposition. The rule by which terms are sorted and grouped needs to be spelled out, and while there is a logical argument for sorting additive terms by absolute differences and not some other measure, it is not necessarily an argument against allowing the user to sort and group the terms by other criteria.

There is another way to interpret the question of which variables have different distributions across the groups, in terms of which variables are conditionally dependent on the group variable, which is essentially asking the question that the difference detection methods answer. A related question that arose when discussing an example is whether there are factors other than those in the parent set that contribute to a particular variable. This suggests that it is important to make the user aware of the assumptions behind the generated explanation. The system first builds a BN model (using either the uni-model or multi-model

method from Chapter 4), a process during which a search of possible parents sets for each variable is performed, and a *maximum-a-posteriori* parent set is selected. The explanation is then generated subject to those parent sets. Both the question of whether factors other than the parents directly influence a variable and the question of whether there is conditionally dependent on the group are both questioning the model construction phase, something that is beyond the scope of generating an explanation with respect to a settled-upon model.

This brings up a wider question of whether we should provide a user with model-questioning tools in addition to the current tool which is designed specifically for model explanation with respect to a settled-upon model. Moreover, if it is important to provide such model-questioning tools, are those tools something that should be integrated into the workflow of difference explanation, or should they be treated as stand-alone applications for editing the 'input' to the difference explanation application?

Providing more analysis tools and capability to question more of the various stages of the computations that lead to the displayed explanation would expand the user's capacity for exploring the data and models. Exploration, however, also has the disadvantage of having a cost in terms of time and mental effort expended by the user. An attractive feature of the textual explanation summaries I provide in the case studies in Chapter 6 is that there is no exploration burden on the reader. Indeed, discussion with the expert also indicated that there is a desire to jump to "the explanation." There is a sense in which exploration may be seen as a burden to be overcome on the way to "the explanation." In settings where users have such a mindset, the solution is not to enable deeper and more detailed exploration, but rather, to attempt to reduce the amount of information the user needs to process to reach a conclusion. The logical course for such settings is the development of an automated generation of an explanation summary that is simple, yet at just the right depth to find the relevant contributing factors. How to do so in an automated manner is an open question. Picking good testing thresholds for a report generator such as DECC goes a long way towards this goal, but what may work as a good threshold for one set of data may not be appropriate for another, necessitating a certain degree of expert involvement.

8.0 CONCLUSIONS AND FUTURE WORK

8.1 CONTRIBUTIONS

This dissertation is concerned with the task of explaining differences across groups. This is a task that people encounter often, not only in the research environment, but also in less formal settings. Existing statistical tools designed specifically for discovering and understanding differences are limited, as discussed in Chapter 3. The purpose of the methods developed in this dissertation is to provide such tools, and the broader aims of this work are to understand what properties such tools should have to be successful and to motivate further development of new approaches to discovering and understanding differences.

Throughout this dissertation, I take the approach that the detection and explanation of differences between two groups should be performed in the context of data-derived models. I formalize this task by taking a probabilistic approach and defining the task as that of detecting parametric differences in models that (a) have clearly defined associations of parameters to variables, and (b) have independent prior distributions over those parameters. Chapter 4 presents a general conceptual framework for difference detection using any such model. Within this framework, I state the difference recovery hypothesis: Given data generated from two models that have similarities and differences between them, we can recover those model differences from observing the data.

This dissertation focuses on applying this framework to BN models with Dirichlet parameter priors. Two novel approaches for difference detection using BNs are developed, the uni-model approach, which reduces to a specialized type of BN structure and parameter learning, and the more general multi-model approach. These approaches to statistical difference detection have some notable properties. They provide a measure of the overall

difference between two groups of data, a measure for each variable of how much the differences in conditional distributions of that variable contribute to the overall difference, and a measure of how much the differences in a subset of the conditional distributions of a variable contribute to the measure for each variable. Additionally, they can be applied with no prior knowledge about the data. Most other statistical methods that can be applied for difference detection do not have all of these properties There are many univariate methods that can measure individual variable differences, but they do not provide an overall score of difference; on the other hand there are various measures of the multivariate distributions, but they do not have a means to represent individual variable contributions.

Chapter 5 describes an empirical evaluation that tests the difference recovery hypothesis in terms of a statistically-significant-difference-detection task on the multi-model and uni-model approaches, and compares them to a regularized logistic regression baseline. Empirical tests show that the multi-model approach is overall the best for detecting statistically significant differences. It demonstrates that the differences between the generating models can be reliably recovered in many cases. There is, however, room for improvement on the multi-model's detection quality (as measured by detection AUC), and there is a need for the development of better methods.

I also performed an empirical evaluation of the quality of the multi-model approach when combined with measures of clinical significance on the detection of clinically significant differences between datasets. This is a narrower test of the difference recovery hypothesis, where the task is to recover only differences that are clinically significant in the generating models. The evaluation revealed the surprising result that the simpler, non-probabilistic clinical significance tests had better detection performance. The performance measured for the probabilistic tests took the form of computing AUCs by varying the threshold of the probabilistic test's inner test, and keeping the probability threshold constant. It is possible that the probabilistic tests would yield better AUCs if both thresholds were varied. A possible avenue for future work is the testing of other measures of clinical significance for this task.

The detection of model parameter differences is essential to correctly modeling the differences between two groups. However, the questions of interest to researchers examining

the data are not always about the differences in the model. Rather, they are about whether inferences derived using the model yield statistically and clinically significant differences. Chapter 6 describes novel methods for explaining a difference in the inference of the marginal probability of a variable, as computed by a model of the similarities and differences. What I define to be an explanation, in this context, is an account of how the elementary model differences, those that where found using the difference detection methods presented in Chapter 4, contribute to the observed difference we seek to explain. The 'insightful explanation' hypothesis is that the explanations produced in this manner can shed light on important relationships in the data and provide useful insights.

For BN models, this translates to tracing the BN inference used to calculate the marginal probability of interest. The methods I developed accomplish that by traversing the BN structure from the node of interest through its ancestry. Additionally, the difference explanation methods focus on only including clinically significant terms in the explanations.

I presented three explanation methods of increasing complexity. The first, CPR, introduced the concept of explanation by decomposition. A marginal term is explained by the joint terms (the joint probabilities of a node and its parents) that contribute to it, and each joint term is explained by a conditional term and a collection of individual parent terms. The second method, EDAPD, introduced the idea of recursion—explanation of each individual parent term in the same way the initial marginal term is explained. The challenge with recursion is that in order for EDAPD's explanation to be coherent, independence between parent terms is required. I defined a class of BN structures, ASCNs, that ensures the independence between parents. I also presented a greedy algorithm for learning ASCNs from data by maximizing a BD score. The third explanation method, DECC, generalizes EDAPD by including a conditioning term which makes it applicable to general BN structures. I presented case studies in which each explanation method is applied to real clinical data. The case studies demonstrated that clinically meaningful relationships can be uncovered by the explanation methods.

The findings demonstrated in the case studies provide support for the 'insightful explanation' hypothesis. The explanations appear to reveal meaningful relationships between variables. A more thorough evaluation of the hypothesis is an open problem for future work.

One would need to establish a standard for comparison against alternatives; currently, no such standard exists. Once such a standard is established, a large-scale evaluation of the methods on a wide range of datasets could be performed.

In addition to developing CPR, EDAPD, and DECC, which statically generate an explanation of the marginal difference in the distribution of a variable, I created an interactive prototype that allows the user to navigate through an explanation. The main advantage of such an interface is that instead of selecting explanation elements based on pre-determined clinical significance tests, it gives the user flexible access to all elements of the explanation. In this manner a researcher has the freedom to guide the explanation and focus on elements found to be insightful in accounting for differences between the two groups.

The prototype developed is a proof-of concept interactive interface to difference explanation. I discussed the strengths and weaknesses of the prototype, and specified some features that would need to be added for developing an interface that a clinical researcher or data analyst could use in their data analysis workflow. An interactive interface can also aid in further addressing the 'insightful explanation' hypothesis. A thorough user study could evaluate the success of the explanation methods at conveying the information about the differences in the data to the users, providing a natural metric for insightfulness.

Together, the difference recovery hypothesis and the 'insightful explanation' hypothesis address the general hypothesis that this thesis set out to explore: *One can systematically produce explanations that are more revealing and insightful than those obtained from traditional methods by approaching the problem of comparing a pair of groups as that of identifying significant local distributional differences between two multivariate distribution estimates for those groups and explaining their effects on variables of interest.* This dissertation developed and explored methods for the systematic generation of explanations that describe differences of interest between a pair of groups in terms of local distributional differences. There was not a thorough evaluation of how insightful such explanations are, but case studies demonstrated that important relationships in the data are revealed in the explanation. It is an open problem to explore further the degree and frequency to which explanations generated in this manner are insightful.

8.2 FUTURE WORK

There were various technical challenges encountered throughout the development of the methods in this dissertation that can be addressed in many ways. In this section I discuss alternate or extended approaches to these challenges, which are beyond the scope of this dissertation work.

8.2.1 Confounding relationships in the uni-model approach

Section 4.2.2 discusses that the uni-model approach to difference detection is designed to detect statistical differences that may or may not be causal differences.

There is a clear path to the development of explanation faculties that explain whether differences are causal or not. One could learn a network without constraints on the placement of Z, compute the network changes that are required to produce a network with an orphaned Z but statistically equivalent relationships, and when generating an explanation using the new network, trace newly found relationships back to their counterparts in the unconstrained network. An explanation produced in this manner would make explicit whether a given relationship is causal (assuming causal sufficiency), or merely statistical due to the fixing of Z within each group or whether it is present as-is in the joint data.

8.2.2 Extension of the multi-model approach to many groups

This dissertation has focused on comparing a pair of data groups. While the pairwise comparison of groups of data is widely applicable, there are many situations in which there are more than two groups to compare. Often in such situation the comparison of interest is how one group (e.g. $\mathcal{D}_1$) differs from the average of the rest of the groups (e.g. $\mathcal{D}_2, \ldots \mathcal{D}_D$). In such cases the current approach is still applicable since the comparison of interest is the binary comparison of $\mathcal{D}_1$ to $\bigcup_{d=2}^{D} \mathcal{D}_d$.

In some cases, however a truly multi-way comparison is of interest, and the methods presented in this dissertation can extend to such cases. The posterior odds that we obtain for pairs of groups in the multi-model approach can be obtained for a larger collection of groups

by a straightforward extension: Suppose that we are given D groups of data, $\mathcal{D}_1, \ldots, \mathcal{D}_D$. We can similarly learn $D+1$ MAP models, $\mathcal{M}_1, \ldots, \mathcal{M}_D$ from each group individually, and $\mathcal{M}_\cup$ from the complete data.

For a variable X_i, we extend the notation to indicate its parent sets in $\mathcal{M}_1, \ldots, \mathcal{M}_D$, and $\mathcal{M}_\cup$ by $\Pi_i^{(1)}, \ldots, \Pi_i^{(D)}$, and $\Pi_i^{(\cup)}$, respectively. Let $\overset{\cap}{\Pi}_i = \left(\bigcap_{d=1}^{D} \Pi_i^{(d)}\right) \cap \Pi_i^{(\cup)}$ and enumerate the configuration of this subset of parents as $\eta = 1, \ldots, H_i$. Then define $T_{i\eta}, T_i, T$ as before, and $S_{i\eta}, S_i, S$ as:

$$S_{i\eta} = \prod_{d=1}^{D} \left(\prod_{j \in J_i^d(\eta)} BD_{ij}(\mathcal{M}_d, \mathcal{D}_d) \right) \tag{8.1}$$

$$S_i = \prod_{\eta=1}^{H_i} S_{i\eta} \tag{8.2}$$

$$S = \prod_{i=1}^{n} S_i. \tag{8.3}$$

Using these extensions to the notation, posterior odds for the presence of a difference can be obtained using (4.24) and (4.23).

This comparison, however, only compares the hypotheses of whether a parameter group $\boldsymbol{\Theta}_{i\eta}$ is different across all groups, or shared among all groups. There are other combinations that may be considered: a set of parameters may be shared among one subset of groups, while another subset of groups shares a different set of parameter. A more comprehensive approach would account for all such possibilities. That means that, for example, if we have three groups of data $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ to compare, we would have to learn six MAP models: $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_{1\cup2}, \mathcal{M}_{1\cup3}, \mathcal{M}_{2,\cup3}$ and $\mathcal{M}_\cup$. We would then have to make a comparison for each parameter group $\boldsymbol{\Theta}_{i\eta}$ between

$$E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_1} P(\mathcal{D}_1|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_1) \times E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_2} P(\mathcal{D}_2|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_2) \times E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_3} P(\mathcal{D}_3|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_3) \ , \tag{8.4}$$

$$E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_{1\cup2}} P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_{1\cup2}) \times E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_3} P(\mathcal{D}_3|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_3) \ , \tag{8.5}$$

$$E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_{1\cup3}} P(\mathcal{D}_1, \mathcal{D}_3|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_{1\cup3}) \times E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_2} P(\mathcal{D}_2|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_2) \ , \tag{8.6}$$

$$E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_1} P(\mathcal{D}_1|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_1) \times E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_{2\cup3}} P(\mathcal{D}_2, \mathcal{D}_3|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_{2\cup3}) \ , \tag{8.7}$$

$$\text{and } E_{\boldsymbol{\Theta}_{i\eta}|\mathcal{M}_\cup} P(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3|\boldsymbol{\Theta}_{i\eta}, \mathcal{M}_\cup) \ . \tag{8.8}$$

Searching over all such possibilities is a nontrivial problem, since the number of possible groupings is exponential in the number of elements to group.

8.2.3 Extending to data with differing variable sets

In this dissertation the focus was on comparing two groups of data that share a single variable set $\mathbf{X}$ where all variables are observed. In some cases we might have differing variable sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ available for $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Such situations may arise when information in each group is collected differently, or different information is collected about the two groups. This may happen, for example, when comparing two hospitals that collect different information about their patients.

When the variable sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ overlap, the multi-model approach may point to a method of comparing these groups. We can construct the MAP models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ over the variable sets $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(1)} \cap \mathbf{X}^{(2)}$ respectively, a posterior odds for a difference in the parameters defining each variable in $\mathbf{X}^{(1)} \cap \mathbf{X}^{(2)}$ can be obtained, and the synthesis process of Section 4.3.4 can be used to build a BN modeling both sets of data. This approach has the effect of assuming that the variables that are not observed in one group have the same conditional distribution in the other group; in some situations this assumption may be a reasonable approach. The result is a model of a joint distribution over $\mathbf{X}^{(1)} \cup \mathbf{X}^{(2)}$, which we may use to generate explanations. In addition to explanation of differences, this constructed model has potential uses as a method for imputation.

8.2.4 Learning multi-model ASCN

An ASCN is a type of BN structure that EDAPD requires for operation. The ASCN learning algorithm I presented in Section 6.2.1 follows the uni-model approach. There is no direct analogy for multi-model learning: we cannot apply the multi-model approach to ASCNs, since, even if $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_\cup$ are SCNs, there is no guarantee that model synthesis would not create loops in the network structure. It would be interesting to develop a method that allows the learning of an ASCN structure, while having the ability of the multi-model to compare variables that have one set of parents in one group and another in the other.

A possible way to approach this is to perform the model learning jointly. Recall that to synthesize a single BN model from $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$, we consider each variable X_i individually, and determine the parents of X_i based on the posterior odds of seeing a difference in X_i generally, as well as the posterior odds that each parameter group $\mathbf{\Theta}_{i\eta}$ differs. All we need to know about the structure of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ to compute these odds is the parent set of X_i in each model. A greedy learner that iterates through the variables in topological order and locks in the parent sets for $X_1, \ldots, X_{i-1}$ before searching for the set of parents for X_i (analogously to the greedy ASCN learner in Section 6.2.1) could constrain the sets of possible parents for X_i in each of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ in such a way as to ensure that the parent set resulting from the synthesis of the models would not violate the ASCN structure constraint.

Other approaches to multi-model ASCN learning are also possible, as are additional approaches to uni-model ASCN learning not explored in this dissertation.

8.2.5 Other approaches to explanation with multi-models

The approach I present of constructing a single BN from the three compared networks in the the multi-model approach allows for running the explanation procedure on the multi-model. It may be possible to combine the explanation method and the multi-model approach in other ways, that do not require the consolidation of the networks into one.

8.2.6 Using context-specific independence in the explanation process

Section 4.3.4 discusses that the multi-net approach to synthesizing a single BN which captures the differences between the distributions created context-specific independence (CSI) in those nodes where statistically significant parameter differences were detected. Boutilier et al. (1996) discuss methods for efficient representation of CSIs as well as methods for exploiting CSIs to increase the efficiency of probabilistic inference. In the current dissertation work, however, the BN built from the multi-model approach does not take advantage of these methods, and is treated as a BN with full conditional probability tables throughout the explanation process.

Another way to conceptualize the synthesized BN is to see the joint states of the parents of X_i that are common to $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ as turning arcs from the other parents on and off. Ideas for BN-like models where the structure itself may be dependent on the value of a variable in the model have been explored before. Bayesian multinets discussed by Geiger and Heckerman (1996) and similarity networks discussed by Heckerman (1990) are examples of such models.

It would be interesting to explore whether any of the specific methods developed for CSIs, multinets, and similarity networks can be applied to the improve on the methods for explanation of differences developed here.

8.2.7 Learning context-specific independence

Because of the potential computational and modeling advantages of local CSI in BNs, the learning of CSI from data has been a topic of research interest (Chickering et al., 1997; Friedman and Goldszmidt, 1998).

The network synthesis method in Section 4.3.4 builds nodes with context-specific independence: the partial parent assignment associated with η determines which remaining parents of the variable will be used in determining the conditional distribution of X_i. Additionally, note that the synthesis takes three models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ over the set of variables $\mathbf{X}$ and produces a single model over the set of variables $\mathbf{X} \cup \{Z\}$.

This suggests that we may view the synthesis method as a process for incorporating a new variable into a model. A logical extension is to consider whether we could add multiple variables in this manner. Indeed, let the notation $\textsc{Synthesized}(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_\cup, Z)$ represent the synthesized single BN containing Z created from $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$, which do not contain Z. Let us suppose that the variables $\mathbf{X} = (X_1, \ldots, X_n)$ are binary, where x_{i1} and x_{i2} are the possible values of each X_i. Let the notation $(\mathcal{D}|x_{ik})$ denote the set of records in $\mathcal{D}$ for which $X_i = x_{ik}$. Consider the following algorithm:

function $\textsc{Learn}(\mathcal{D}, (X_i, \ldots, X_n))$ ▷ Learn a BN over $X_i, \ldots, X_n$.
  if $i = n$ **then**
    return a model with a single node X_n

else

$\mathcal{M}_1 \leftarrow \textsc{Learn}((\mathcal{D}|x_{i1}), (X_{i-1}, \ldots, X_n))$

$\mathcal{M}_2 \leftarrow \textsc{Learn}((\mathcal{D}|x_{i2}), (X_{i-1}, \ldots, X_n))$

$\mathcal{M}_\cup \leftarrow \textsc{Learn}(\mathcal{D}, (X_{i-1}, \ldots, X_n)))$

return $\textsc{Synthesized}(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_\cup, X_i)$

Calling $\textsc{Learn}(\mathcal{D}, \mathbf{X})$ would then learn a full BN where each node has the sort of context-specific independence with respect to all of its parents that each node in the synthesized BN from Section 4.3.4 has with respect to Z.

It is easy to see that the naïve implementation above has time complexity that is exponential in the number of variables. It is possible that with the aid of techniques such as branching and bounding the search space or dynamic programming, a more efficient approach can be developed. Additionally, it may be possible to develop heuristic methods that use the same basic network construction principles, but improve on the computational complexity by sacrificing optimality.

8.2.8 Extending explanation to complex inferences

The explanation methods presented here have all focused on providing an account of how a difference in the marginal distribution of a single variable is derived from the elemental differences in the model. It is possible, however, that a researcher may be interested in a more complex query, the most general form of which is an account of why $P(\mathbf{a}|\mathbf{e}, Z = 1)$ differs from $P(\mathbf{a}|\mathbf{e}, Z = 2)$ according to the model. In principle, DECC may be applied to such a query. For $\mathbf{a} = (a_1, \ldots, a_m)$, we have the decomposition

$$P(\mathbf{a}|\mathbf{e}, z) = \prod_{t=1}^{m} P(a_t|a_1, \ldots, a_{t-1}, \mathbf{e}, z) \tag{8.9}$$

in which each term on the right-hand side can act as input to DECC.

Potential future research would include the evaluation of how well such explanations perform and the development of extensions for addressing any challenges that arise in such application.

8.2.9 Explanation using a factor tree

Section 6.3 discusses the problem that loops in the undirected BN structure introduce. Loops may create dependence between parents, which would create inconsistency in an explanation generated by EDAPD, since EDAPD explains the contribution of the joint ratio term as a product of the conditional ratio term and a product of marginal parent terms. I discussed the idea of using cut-set conditioning to address the problem and pointed out the challenges with this approach. Finally, a different approach where conditioning terms are carried with the explanation is used in DECC.

Another approach to tackling the loop problem in BN inference is the conversion of the BN into a factor tree. Lauritzen and Spiegelhalter (1988) present a method whereby a BN is converted into a tree of factors, where each of the factors is a group of original BN variables. The multivariate distribution represented by the BN can instead be represented as a product of value assignments of those factors. The advantage of switching to this representation is that inference can then be performed by straightforward belief propagation over the factor tree.

It may be possible to apply the same approach to explanation: instead of traversing a BN structure, the traversal happens over factors, and differences are explained in terms of numerical contributions from differences in the values of the factors across groups. Such an approach would ensure a consistent explanation of differences throughout explanation steps and have a direct correspondence to the inference process. Moreover, such an explanation would be EDAPD-like because the factor tree is singly-connected. The disadvantage of this approach is that directionality is absent in the factor tree, and the explanation would not be necessarily directed "upwards" in the ancestry, potentially making proper interpretation challenging. Another disadvantage is that factors can potentially represent the joint probabilities of many variables, making the contribution of any one factor difficult to interpret.

8.2.10 Adding statistical significance testing during explanation

Throughout the explanation generation processes, we test marginal probabilities and joint probabilities for clinical significance. In contrast to this, only the conditional distributions—

the BN parameters—are tested for statistical significance. When a difference or ratio of probabilities is reported to the user, when it is comparing the conditional probability of a variable given its parents, it corresponds to a BN parameter and is therefore known to have been tested for statistical significance. When the difference or ratio is of another sort, such as that of joint probabilities, or marginal probabilities that have parents other than Z in the explanation BN, the probabilities used are a result of inference. While the inference is based on a BN in which all parameters have been tested for statistical significance, the difference or ratio of the probability of interest has only been directly tested for clinical significance.

Hence, it may be of interest to add tests of statistical significance throughout the explanation process. For example, for each probability, difference of probabilities, and ratio of probabilities that is reported we can estimate a Bayesian credibility interval by sampling network parameters from their posterior Dirichlet distributions, the same process used in the probabilistic clinical significance tests of Section 4.4. Another, complementary approach, is to check the probabilities obtained from inferences against the proportions obtained by checking the corresponding counts in the data. This has the added benefit of checking how well the model fits the data, particularly in regard to the query of interest.

Some questions to explore in this line of research are: how commonly are statistically insignificant marginal and joint probabilities found clinically significant by the existing tests, how much explanations are improved by various statistical significance filters, whether such filters can help understand an explanation better, and whether the presence and statement of statistical significance tests affect a user's trust in the explanation provided.

8.2.11 Model averaging and ensemble models

In this dissertation, a model that codifies the differences between the distributions of the two groups is used to compute the probabilities used throughout the explanation process; namely, the network containing Z produced either directly from learning in the uni-model approach, or by construction based on MAP models in the multi-model approach. Section 1.2 briefly discusses the potential pitfalls of committing to a single model. If the model does not accurately represent the data, the probabilities that result from inference may not match

the proportions of counts in the data, and in the extreme worst case this mismatch may lead to incorrect conclusions. Moreover, committing to a single model that dictates a single explanation leaves the user blind to potential alternative explanations that may be as good or almost as good at explaining the differences observed.

The idea of adding statistical testing at the explanation level in Section 8.2.10 is one approach to mitigating the risk of drawing incorrect conclusions from an inaccurate model. An approach that can improve the Bayesian statistical significance tests (and credibility intervals) is that of averaging not only over model parameters, but also over model structures. There are multiple ways to define the space of structures to average. A simple averaging can be done over the two possibilities of including or excluding Z as a parent of each node; more general averaging can be done by averaging over all possible parent sets of each node subject to a node ordering; or the most general averaging can be done by averaging over all structures and orderings. The main challenge in the latter two approaches is to find a way to handle (both in terms of computational complexity and explanation clarity) the large space over which to average. One approach is to approximate the average by only averaging over a few of the most probable structures.

Model averaging can be used not only as a tool for statistical significance testing, but also as an alternative means to guide the explanation. This is one way to move away from the commitment to a single model, since the averaging reflects conclusions based on all models in the space over which we average. It is possible to develop an explanation that would give the user a distribution over possible parents and present distributions of contributions to a difference in each step of the explanation. A challenge in developing such a method is to define what quantities should be presented at each step and how. An observed difference is the result of potentially different divisions of contributions between different terms for models with different parent sets. To average all these contributions and associate them with a distribution over possible parent sets, a novel approach to appropriately aggregating and presenting the information would need to be developed.

Another means of moving away from the commitment to a single model is to look at an ensemble of best models. Instead of averaging the information from many models to obtain a single answer and a single explanation, we would look at the ensemble as a collection

of alternatives. These models and their corresponding explanations can be ranked by how well they fit the data, by the complexity or size of the explanation. We could also search for patterns and relationships that are common to many of the models and explanations, and report those (since we expect a pattern that reliably reappears in various models to indicate a strong relationship in the data). Considering an ensemble of models also has the advantage of allowing us to present a user with multiple alternative explanations. Having multiple explanations can be very useful: it can guide further investigation and help present a more complete picture of the space of likely possibilities.

8.3 CONCLUDING REMARKS

This dissertation presented a novel approach to comparing groups of data points. The process of comparing groups of data was divided into multiple stages: The learning of MAP models for the data in each group, the identification of statistical differences between model parameters, the construction of a single model that captures those differences, and finally, the explanation of inferences of differences in marginal distributions in the form of an account of clinically significant contributions of elemental model differences to the marginal difference. A general framework for the process was presented, and while this dissertation focuses on BNs over multinomial variables, the framework provides guidelines for its application to a broad range of model types.

The methods for detecting statistical differences in parameters, particularly the multi-model approach which computes the posterior odds of parameter differences at multiple levels, has a combination of properties that to my knowledge no other statistical methods for comparing groups have. It measures whether the group distributions differ overall, clearly expresses how differences at the variable and parameter-group level contribute to the measure, and requires no prior knowledge for application. This motivates further development of statistical methods that can fulfill the same combination of properties. The method presented for constructing a single BN model based on the multi-model approach builds a network with local context-specific independence, which can also be viewed as a

model where the graphical structure is dependent on the values of variables, in a manner akin to Bayesian multinets and similarity networks. The BN synthesis method suggests a novel approach to learning BNs with context-specific independence from data, and may have similar implications for learning Bayesian multinets or similarity networks from data.

This dissertation presented methods for the explanation of differences captured by a BN. The methods focused on explaining the differences in a marginal probability distribution that was computed from the model by inference. Case studies on clinical data demonstrated that the methods revealed medically sensible accounts of the differences of the differences observed in the data. There are many possible aspects of the explanation methods that can be extended in interesting and useful ways, such as the explicit representation of causal and non-causal relationships in the models, explicit statistical testing at various stages of the explanation procedure, and the leveraging of context-specific independencies in the model. Other interesting avenues for future work are the extension of explanation to explaining differences in more general inferences, and the application of explanation-by-traversal to models other than BNs, for example, to the factor tree of a BN.

A prototype analysis tool that implements the detection and explanation methods was developed and implemented. The discussion of the design considerations throughout the development process, as well as the strengths and weaknesses of the prototype and of desired features for future development, are intended to help future researchers in developing similar data analysis tools for performing comparisons between groups.

In addition to the direct contribution of the research presented of providing methods for detecting and explaining differences between groups of data, the research opens many avenues for future work, both in the form of extensions of and improvements on the method, and in the form of discoveries relevant to related topics.

BIBLIOGRAPHY

S. Acid and L.M. De Campos. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 3–10. Morgan Kaufmann Publishers Inc., 1996.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Stephen D Bay and Michael J Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

DA Bell, W Liu, J Cheng, R Greiner, and J Kelly. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.

R. Bouckaert. Probabilistic network construction using the minimum description length principle. *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 41–48, 1993.

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 115–123. Morgan Kaufmann Publishers Inc., 1996.

George EP Box and Norman R Draper. *Empirical model-building and response surfaces.* John Wiley & Sons, 1987.

Eugene Charniak and Solomon Eyal Shimony. Cost-based abduction and map explanation. *Artificial Intelligence*, 66:345–374, 1994. doi: 10.1016/0004-3702(94)90030-2.

X.W. Chen, G. Anantha, and X. Lin. Improving bayesian network structure learning with mutual information-based node ordering in the k2 algorithm. *Knowledge and Data Engineering, IEEE Transactions on*, 20(5):628–640, 2008.

Jie Cheng, David A Bell, and Weiru Liu. An algorithm for bayesian belief network construction from data. In *proceedings of AI & STAT97*, pages 83–90, 1997.

D. M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag, 1996.

D. M. Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

D. M. Chickering and C. Meek. On the incompatibility of faithfulness and monotone dag faithfulness. *Artificial Intelligence*, 170(8):653–666, 2006.

David Maxwell Chickering, David Heckerman, and Christopher Meek. A bayesian approach to learning bayesian networks with local structure. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 80–89. Morgan Kaufmann Publishers Inc., 1997.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462 – 467, may 1968. ISSN 0018-9448. doi: 10.1109/TIT.1968.1054142.

Ton J Cleophas and Aeilko H Zwinderman. Post-hoc analyses in clinical trials, a case for logistic regression analysis. *Statistics Applied to Clinical Studies*, pages 227–231, 2012.

Frederick L Coolidge. *Statistics: A gentle introduction*. Sage Publications, 2012.

G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992. URL http://www.springerlink.com/index/10.1007/BF00994110.

R. Daly, Q. Shen, and S. Aitken. Learning Bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, 26(2):99–157, 2011.

L.M. de Campos and J.F. Huete. Approximating causal orderings for bayesian networks using genetic algorithms and simulated annealing. In *Proceedings of the Eight Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 333–340, 2000a.

L.M. De Campos, JM Puerta, et al. Learning bayesian networks by ant colony optimisation: searching in two different spaces. *Mathware & soft computing*, 9(3):251–268, 2008.

Luis M de Campos. Independency relationships and learning algorithms for singly connected networks. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(4):511–549, 1998.

Luis M de Campos and Juan F Huete. On the use of independence relationships for learning simplified belief networks. *International Journal of Intelligent Systems*, 12(7):495–522, 1997.

Luis M de Campos and Juan F Huete. A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning*, 24(1):11–37, 2000b.

M. L. de Campos, M. J. Fernández-Luna, and M. J. Puerta. Local search methods for learning Bayesian networks using a modified neighborhood in the space of DAGs. *Advances in Artificial IntelligenceIBERAMIA 2002*, pages 182–192, 2002.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, pages 837–845, 1988.

Marek Jozef Druzdzel. *Probabilistic reasoning in decision support systems: from computation to common sense*. PhD thesis, Pittsburgh, PA, USA, 1993. UMI Order No. GAX93-22863.

Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31:1–38, 2004.

Nir Friedman and Moises Goldszmidt. Learning bayesian networks with local structure. In *Learning in graphical models*, pages 421–459. Springer, 1998.

Nir Friedman, Dan Geiger, Moises Goldszmidt, G. Provan, P. Langley, and P. Smyth. Bayesian network classifiers. In *Machine Learning*, pages 131–163, 1997.

José A. Gámez. Abductive inference in bayesian networks: A review. In Antonio Salmerón José A. Gámez, Serafín Moral, editor, *Advances in Bayesian Networks*, pages 101–117. Springer, Berlin, 2004.

D. Geiger, A. Paz, and J. Pearl. Learning causal trees from dependence information. In *Proceedings of the Eith National Conference on Artificial Intelligence (AAAI 1990)*, pages 770–776. AAAI Press, 1990a.

Dan Geiger and David Heckerman. Knowledge representation and inference in similarity networks and bayesian multinets. *Artificial Intelligence*, 82(1):45–74, 1996.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990b.

Clark Glymour, Peter Spirtes, and Richard Scheines. Causal inference. In *Erkenntnis Orientated: A Centennial Volume for Rudolf Carnap and Hans Reichenbach*, pages 151–189. Springer, 1991.

Kui Xiang Gou, Gong Xiu Jun, and Zheng Zhao. Learning bayesian network structure from distributed homogeneous data. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on*, volume 3, pages 250–254. IEEE, 2007.

D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-60032-3. URL http://portal.acm.org/citation.cfm?id=308574.308676.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995. ISSN 0885-6125. URL http://dx.doi.org/10.1023/A:1022623210503.

David Heckerman. Probabilistic similarity networks. *Networks*, 20(5):607–636, 1990.

K. B. Hwang, J. Lee, S. W. Chung, and B. T. Zhang. Construction of large-scale Bayesian networks by local to global search. *PRICAI 2002: Trends in Artificial Intelligence*, pages 501–518, 2002.

Harold Jeffreys. *The theory of probability*. Oxford University Press, 1998.

Norman L Johnson, Samuel Kotz, and N Balakrishnan. *Continuous Multivariate Distributions, volume 1, Models and Applications*, volume 59. New York: John Wiley & Sons, 2002.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

Wishwa N Kapoor. *Assessment of the Variation and Outcomes of Pneumonia: Pneumonia Patient Outcomes Research Team (PORT) Final Report*. Agency for Health Policy and Research (AHCPR), 1996.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Alan E Kazdin. The meanings and measurement of clinical significance. 1999.

Ken Kelley and Kristopher J Preacher. On effect size. *Psychological methods*, 17(2):137, 2012.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Carmen Lacave and Francisco J. Diez. A review of explanation methods for bayesian networks. *Knowledge Engineering Review*, 17:2002, 2000.

Carmen Lacave, Manuel Luque, and Francisco Javier Díez. Explanation of bayesian networks and influence diagrams in elvira. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(4):952–965, 2007.

P. Larrañaga, M. Poza, Y. Yurramendi, R.H. Murga, and C.M.H. Kuijpers. Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(9):912–926, 1996.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

F. Liu and Q. Zhu. The max-relevance and min-redundancy greedy Bayesian network learning algorithm. *Bio-inspired Modeling of Cognitive Tasks*, pages 346–356, 2007.

F. Liu, F. Tian, and Q. Zhu. An improved greedy Bayesian network learning algorithm on limited data. *Artificial Neural Networks–ICANN 2007*, pages 49–57, 2007.

David Madigan, Krzysztof Mosurski, and Russell G Almond. Graphical explanation in belief networks. *In Journal of Computational and Graphical Statistics*, 6:160–181, 1997.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems 12*.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Uncertainty in Artificial Intelligence*, volume 11, pages 403–410, 1995.

J.D. Nielsen, T. Kočka, and J.M. Pena. On local optima in learning bayesian networks. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 435–442. Morgan Kaufmann Publishers Inc., 2002.

Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10:377–403, 2009.

M. J Pazzani. Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(3):416–432, 1991.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D. S. Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual explanation of evidence in additive classifiers. In *Proceedings of the 18th conference on Innovative applications of artificial intelligence - Volume 2*, IAAI'06, pages 1822–1829. AAAI Press, 2006. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597122.1597143.

G. Rebane and J. Pearl. *The recovery of causal poly-trees from statistical data*. UCLA, Computer Science Department, 1987.

Solomon E Shimony. A probabilistic framework for explanation. Technical report, Providence, RI, USA, 1991.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter and T. Richardson, editors, *Proceedings of the Twenty-second Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452. AUAI Press, 2006.

A. P. Singh and A. W. Moore. Finding optimal Bayesian networks by dynamic programming. *Techinical report CMU-CALD-05-106*, 2005.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Henri J. Suermondt and Gregory F. Cooper. An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, 26(3):242 – 254, 1993. ISSN 0010-4809. doi: DOI:10.1006/cbmr.1993.1017. URL http://www.sciencedirect.com/science/article/pii/S0010480983710177.

Henri Jacques Suermondt. *Explanation in Bayesian belief networks*. PhD thesis, Stanford, CA, USA, 1992. UMI Order No. GAX92-21673.

Yuriy Sverchkov, Shyam Visweswaran, Gilles Clermont, Milos Hauskrecht, and Gregory F. Cooper. A multivariate probabilistic method for comparing two clinical datasets. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 795–800, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0781-9. doi: 10.1145/2110363.2110460. URL http://doi.acm.org/10.1145/2110363.2110460.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI-05)*. AUAI Press, 2012.

T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth international conference on uncertainty in artificial intelligence*, pages 323–330. Morgan Kaufmann Publishers Inc., 1992.

C. Wallace and K. Korb. Learning linear causal model by mml sampling. In A. Gammerman, editor, *Causal Models and Intelligent Data Management*, pages 89–111. Springer, 1997.

Geoffrey I Webb, Shane Butler, and Douglas Newlands. On detecting differences between groups. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265. ACM, 2003.

M.L. Wong and K.S. Leung. An efficient data mining method for learning bayesian networks using an evolutionary algorithm-based hybrid approach. *Evolutionary Computation, IEEE Transactions on*, 8(4):378–404, 2004.

M.L. Wong, W. Lam, and K.S. Leung. Using evolutionary programming and minimum description length principle for data mining of bayesian networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(2):174–178, 1999.

C. Yuan, B. Malone, and X. Wu. Learning optimal Bayesian networks using A* search. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 2186–2191, Helsinki, Finland, 2011.